GROUP-LOCAL ANCHORING FOR SUBGROUP INTERVALS IN PROGRAMMATIC WEAK SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Programmatic weak supervision uses approximate labeling functions to construct training labels. Evaluating these labels requires intervals for class proportions within subgroups of an unlabeled pool. An existing approach derives such intervals from pool-wide statistics estimated using a small labeled sample. These global constraints can provide simultaneous coverage while leaving small-group intervals equal to the full probability range, because changes within a small group have little effect on pool-wide statistics. We study how labels reaching a subgroup determine interval precision. For binary labels, we derive a width lower bound from the pool-to-group size ratio and the room for variation allowed by the global constraints. This identifies a computable group-size threshold below which intervals must span the full probability range. Motivated by this bound, group-local anchoring adds constraints estimated from labeled observations within each group. Joint calibration preserves simultaneous coverage when group selection is independent of the estimation labels. For groups learned from labels, we use disjoint samples for selection and estimation. With 400 labeled draws from the pool, median global-to-anchored width ratios range from 1.21 to 1.69 across five text corpora, with gains concentrated in groups containing at least 32 pool instances. Experiments on text and image corpora further show that reusing selection labels for estimation produces coverage failures concentrated in selectors that memorize labels for individual labeling-function patterns. In the image experiments, independent estimation reduces this undercoverage. Median split-to-reference width ratios are 1.42–1.98, where the reference uses all acquired labels on the same groups. These results link subgroup precision to local labeling information and support independent estimation for learned groups.

1 INTRODUCTION

Programmatic weak supervision uses inexpensive labeling functions to construct training labels (Zhang et al., 2022). Assessing these labels often requires the class proportion within a subgroup of the data. For example, when a keyword rule assigns the positive class whenever it fires, the positive-label proportion of its firing group measures the rule's precision. Reliable subgroup estimates can therefore guide which weak labels to use for training.

Label models combine conflicting labeling-function outputs into per-instance probabilities (Ratner et al., 2016; 2017). Partial identification instead describes the range of target values compatible with the available information (Manski, 2003; Maia Polo et al., 2024). For subgroup proportions, Álvarez et al. (2025) use pool-wide statistics estimated from a small labeled sample to constrain candidate label distributions. Minimizing and maximizing a group's proportion over those distributions produces an interval. If the constraints contain the true distribution, intervals for all subgroups cover their targets simultaneously.

This coverage property leaves a question of interval width. Changing labels within a small group can greatly alter its own class proportion while barely affecting pool-wide statistics. A keyword rule that fires on a small subset of reviews illustrates the issue: the global constraints may accommodate both mostly positive and mostly negative labels in that subset. An interval can consequently cover the truth while spanning the entire probability range. We study when global information forces such wide intervals and how labels sampled within a group can improve their precision.

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.

Our analysis first quantifies the limitation of global constraints in the binary-label setting. For the nonempty feasible sets and bounded statistics considered, we derive a width lower bound governed by the pool-to-group size ratio and the *center slack*, the largest attainable minimum margin to the global constraint boundaries. The bound gives a computable sufficient condition for a group's interval to span $[0, 1]$ (Theorem 3). This result motivates *group-local anchoring*: use the labeled observations that fall within a group to constrain its proportion directly, alongside the global statistics.

For groups chosen independently of the estimation labels, local binomial confidence bounds provide these additional constraints. Calibrating the global and local constraints jointly preserves simultaneous coverage over all queried subgroups (Section 4). The number of sampled labels reaching a group determines how much its anchor can narrow the interval. With enough local observations and proportions away from the boundaries, the variance-dependent term decreases with the square root of the local sample size. At a fixed pool-wide sampling budget, this gives a square-root dependence on the pool-to-group size ratio, in the regime where that term dominates the finite-sample correction.

Groups learned from labels introduce an additional selection effect. A selector can favor groups that look unusually pure on the same observations later used to estimate their proportions. We address this dependence by dividing labeled draws between group selection and constraint estimation. Conditional on the selection sample, the groups are fixed and the other sample still supplies independent uniform draws for estimation. This construction yields simultaneous coverage without restricting the selector's complexity (Section 5).

Experiments examine interval width and label reuse separately. Across five text corpora with 400 labeled draws, groups containing at least 32 instances have corpus-level median global-to-anchored width ratios of 1.77–2.41 (Table A12). Thus the gains concentrate where enough observations reach the groups. For learned text groups, independent estimation reduces the unshrunk pattern selector's simultaneous-coverage failures from 156 to 3 among 640 domain–repetition units, where a unit fails if any queried group is uncovered (Section 6.4). These results connect interval precision to local information and coverage to how that information is used.

Our contributions are as follows.

- We establish a lower bound on subgroup interval width under global constraints and a computable sufficient condition for full-range intervals (Theorem 3).
- We introduce group-local anchoring with joint coverage control and characterize how local sample size governs attainable precision (Section 4).
- We extend the construction to groups learned from labels through independent selection and estimation, and evaluate both width gains and protection against label reuse (Sections 5 and 6).

2 Related Work

Group-conditional guarantees. Conformal prediction covers future outcomes (Vovk et al., 2005). Multiaccuracy controls average errors within subgroups (Kim et al., 2019) and multicalibration additionally conditions on predicted scores (Hebert-Johnson et al., 2018), while conditional conformal methods (Gibbs et al., 2023) and multivalid prediction (Jung et al., 2023) extend prediction-set guarantees across specified groups so an overall coverage rate does not conceal differences between populations. We estimate label proportions in a fixed pool and study the precision of intervals that cover all subgroup proportions simultaneously.

Programmatic weak supervision. Data programming (Ratner et al., 2016) and Snorkel (Ratner et al., 2017; 2019; Fu et al., 2020) combine conflicting labeling-function outputs into probabilistic training labels. Subsequent approaches accommodate dependent sources (Mazzetto et al., 2021) and learn weakly supervised predictors end to end (Cachay et al., 2021), allowing weak signals to inform learning when simple independence assumptions are unsuitable. Weak-supervision evaluation also uses partial identification (Maia Polo et al., 2024) and predictive sets (Cauchois et al., 2022). We build on the confidence-constraint construction of Álvarez et al. (2025), show that its width saturates at the full range below a computable group size, and add local rows whose width the intervals inherit.

Simultaneous and selective inference. Bonferroni correction and Holm's refinement control errors across collections of decisions (Holm, 1979). For data-dependent targets, post-selection inference uses simultaneous protection (Berk et al., 2013) or separates selection from estimation (Wasserman & Roeder, 2009). We apply these principles to global constraints and Clopper–Pearson group anchors (Clopper & Pearson, 1934), allocating confidence jointly while using independent estimation labels to preserve the sampling relation for learned groups.

Variance-sensitive concentration. Concentration inequalities translate bounded observations into confidence bounds for unknown means (Hoeffding, 1963). Empirical Bernstein bounds use sample variance (Maurer & Pontil, 2009) to adapt to the spread of the data. We use this form to bound the Clopper–Pearson radius in terms of the number of true labels that we see in each group. Thus, local labels directly curb the loss of precision caused by constraints on the whole pool.

Label-efficient estimation. Stratified sampling (Neyman, 1934) and active testing (Kossen et al., 2021) allocate labels to improve estimation under a limited budget. Prediction-powered inference (PPI) corrects a prediction-based estimate using labeled residuals (Angelopoulos et al., 2023a). This correction connects uncertainty to prediction errors within the population of interest. PPI variants tune prediction weights (Angelopoulos et al., 2023b), shrink estimates across tasks (Li & Ignatiadis, 2025) and enable sequential evaluation (Csillag et al., 2025; Park et al., 2025). StratPPI adapts correction and allocation across strata (Fisch et al., 2024) to account for heterogeneous prediction errors across subpopulations. We study overlapping queries in a finite pool under one shared label budget. LLM judges support aggregate estimates and risk tests (Lee et al., 2025; Chen et al., 2026; Feng et al., 2026). Local anchoring instead constrains finite-pool subgroup proportions using binomial intervals. Active statistical inference prioritizes informative labels (Zrnic & Candès, 2024). Robust active sampling accounts for inaccurate uncertainty scores (Li et al., 2025), whereas we study subgroup resolution under uniform sampling and independent estimation after selection.

3 Global constraints and subgroup resolution

We define the subgroup target and the global constraints that give simultaneous coverage. We then relate interval width to the variation these constraints allow within a group.

Setting and simultaneous coverage. To assess weak labels within a subgroup, we estimate its label proportion from a small labeled sample of the pool. The pool contains n instances with unknown labels $y_i \in \mathcal{Y} = \{1, \dots, T\}$. Here T is the number of classes and i indexes a pool instance. We condition on this finite pool, its labels and its labeling-function outputs; confidence probabilities refer to the random labeled draws. We observe the outputs of m labeling functions for every instance. The vector Λ_i collects these outputs, each of which is a label or an abstention or a score. For any non-empty group $I \subseteq \{1, \dots, n\}$, we seek its label proportion $p_I^\star(y) = \frac{1}{|I|} \sum_{i \in I} \mathbb{1}\{y_i = y\}$, where $|I|$ is the group size and $\mathbb{1}\{y_i = y\}$ indicates whether instance i has label y. We obtain labels at positions $J = (J_1, \dots, J_\ell)$ drawn independently and uniformly with replacement from the pool. The budget ℓ counts draws, including repeated pool positions.

Global confidence constraints. To relate the labeling-function outputs to the unknown labels, Álvarez et al. (2025) use bounded features $\mathbf{\Phi}(\Lambda_i, y) \in [0,1]^d$. Here d is the feature dimension. The pool moment $\boldsymbol{\tau} = \frac{1}{n} \sum_i \mathbf{\Phi}(\Lambda_i, y_i)$ averages these features at the true labels. For example, $\Phi_j(\Lambda_i, y) = \mathbb{1}\{\Lambda_{ij} = y\}$ records agreement with labeling function j. The labeled draws give $\hat{\boldsymbol{\tau}} = \ell^{-1} \sum_{r=1}^{\ell} \mathbf{\Phi}(\Lambda_{J_r}, y_{J_r})$. A tolerance vector $\boldsymbol{\lambda} \succeq 0$ specifies the allowed estimation error in each coordinate. These estimates constrain candidate assignments of label probabilities to pool instances. Let p_i be a candidate probability distribution on $\mathcal{Y}$ and let $p = (p_1, \dots, p_n)$ collect these distributions. These variables enumerate compatible assignments; the fixed truth is represented by $p_i^\circ(y) = \mathbb{1}\{y_i = y\}$. The set of assignments compatible with the global estimates is

$$\mathcal{U} = \Big\{ p \in \Delta(\mathcal{Y})^n : \big| \tfrac{1}{n} \textstyle\sum_i \mathbb{E}_{p_i} \mathbf{\Phi}(\Lambda_i, y) - \hat{\boldsymbol{\tau}} \big| \preceq \boldsymbol{\lambda} \Big\}, \tag{1}$$

where $\Delta(\mathcal{Y})$ is the probability simplex and $\mathbb{E}_{p_i}$ averages over a candidate label drawn from p_i. The absolute value and inequality act coordinatewise. We obtain the interval for this

group proportion by minimizing and maximizing its candidate value over $\mathcal{U}$. The endpoints are $\underline{p}_I(y) = \min_{p\in\mathcal{U}} \frac{1}{|I|}\sum_{i\in I} p_i(y)$ and $\overline{p}_I(y) = \max_{p\in\mathcal{U}} \frac{1}{|I|}\sum_{i\in I} p_i(y)$. Each endpoint is a linear program, with one simplex per instance, or one probability variable per instance for binary labels. We denote their difference by $W_I(y)$ and consider endpoints whenever the feasible set is non-empty. The following result explains why one confidence event suffices for every group. Here $\delta \in (0,1)$ is the allowed probability of simultaneous coverage failure.

Proposition 1 (Simultaneous coverage, Álvarez et al., 2025, Thm. 1)**.** *If $\mathcal{U}$ is non-empty and $\mathbb{P}\{|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\} \geq 1-\delta$, then with probability at least $1-\delta$ the intervals contain $p^\star_I(y)$ simultaneously for all non-empty $I \subseteq \{1,\dots,n\}$ and all $y \in \mathcal{Y}$.*

The failure budget protects the d constraints jointly. Every queried subgroup shares this truth-inclusion event, so querying additional groups requires no further allocation. The independent-coordinate coverage example and comparison of estimands are in Appendix H.

Width imposed by global constraints. To explain wide intervals for small groups, we measure how far their label probabilities can move without violating a global constraint and write $\rho_I = n/|I| \geq 1$ for the pool-to-group size ratio. We specialize to binary labels $\mathcal{Y} = \{0,1\}$ for the remaining analysis. Write $g(p) = \frac{1}{n}\sum_i \mathbb{E}_{p_i}\boldsymbol{\Phi}(\Lambda_i, y)$ for the pool moment induced by a candidate assignment. We use subscripts k for coordinates of g and of the estimate and tolerance vectors. The box is centered at the estimated moment vector $\hat{\boldsymbol{\tau}}$. We seek a feasible moment whose smallest remaining margin to any box boundary is largest. **Definition 2 (Center slack).** For $p \in \mathcal{U}$, let $s_k(p) := \lambda_k - |g_k(p) - \hat{\tau}_k| \geq 0$ be the remaining margin in coordinate k. The center slack is $s^\star := \max_{p\in\mathcal{U}} \min_{1\leq k\leq d} s_k(p)$.

Theorem 3 (Width lower bound)**.** *Let $\mathcal{U}$ be non-empty and $\Phi_k(\Lambda_i, y) \in [0,1]$ for all i and k and $y \in \mathcal{Y}$. Then every non-empty $I \subseteq \{1,\dots,n\}$ and every $y \in \{0,1\}$ satisfy*

$$W_I(y) \geq \min\{1,\ \rho_I s^\star\}, \qquad \rho_I = n/|I|, \tag{2}$$

where $s^\star$ is the center slack of Definition 2. If $\hat{\boldsymbol{\tau}}$ is realizable, meaning $g(p) = \hat{\boldsymbol{\tau}}$ for some $p \in \Delta(\mathcal{Y})^n$, then $s^\star = \min_k \lambda_k$.

Appendix A proves Theorem 3 by perturbing group probabilities toward the binary extremes. The slack is computable by maximizing an auxiliary $s \geq 0$ subject to $-\lambda_k + s \leq g_k(p) - \hat{\tau}_k \leq \lambda_k - s$ for every k, with the same simplex constraints. Thus $|I| \leq n s^\star$ is a sufficient condition for full-range endpoints under this particular global uncertainty set. For a realizable center and $\lambda_k = \Theta(\ell^{-1/2})$, the bound grows linearly in ρ_I until saturation. Write $\lambda_{\min} = \min_k \lambda_k$ for the minimum global tolerance. For a labeling function that votes for a single class or abstains, a matching upper bound holds on its firing group, establishing a tight group-size exponent in that setting (Appendix A). The binary argument also leads to multiclass bounds using class-specific sensitivity and feasible perturbation directions, with the extensions and numerical comparisons in Appendix E. This width dependence motivates constraining group averages directly so their tolerances do not acquire the pool-to-group factor.

4 Group-local anchoring

We add local confidence constraints at the same scale as the subgroup target and retain the existing endpoint optimization. Their joint validity protects coverage while the labels reaching each group determine precision.

Adding local constraints. We begin with groups whose membership is determined without the labels used to estimate their constraints. **Definition 4 (Admissible family).** A family $\mathcal{I}$ is *admissible* if every $I \in \mathcal{I}$ is a deterministic function of $(\Lambda_1,\dots,\Lambda_n)$ with $|I| \geq 1$. For example, the firing groups of keyword labeling functions form an admissible family. For such a group I, our local target is its positive-label proportion $\tau_I = p^\star_I(1)$. A local confidence constraint, or *anchor*, requires its candidate value $|I|^{-1}\sum_{i\in I} p_i(1)$ to lie in $[\hat{\tau}_I - \lambda_I, \hat{\tau}_I + \lambda_I] \cap [0,1]$. This scalar constraint contributes two inequalities to the endpoint programs.

The same construction permits a vector of local features $\boldsymbol{\Phi}^{\mathrm{loc}}$, with target $\boldsymbol{\tau}_I = |I|^{-1}\sum_{i\in I}\boldsymbol{\Phi}^{\mathrm{loc}}(\Lambda_i, y_i)$, estimate $\hat{\boldsymbol{\tau}}_I$ and tolerance $\boldsymbol{\lambda}_I$. Our prevalence anchor uses $\Phi^{\mathrm{loc}}(\Lambda_i, y) =$

$\mathbb{1}\{y = 1\}$. **Definition 5 (Group-locally anchored uncertainty set).** For an admissible anchor family $\mathcal{I}$, let

$$\mathcal{U}^{\mathrm{aug}} := \mathcal{U} \cap \bigcap_{I \in \mathcal{I}} \Big\{ p : \big| \tfrac{1}{|I|} \textstyle\sum_{i \in I} \mathbb{E}_{p_i} \mathbf{\Phi}^{\mathrm{loc}}(\Lambda_i, \cdot) - \hat{\boldsymbol{\tau}}_I \big| \preceq \boldsymbol{\lambda}_I \Big\}. \tag{3}$$

For any queried group Q, minimizing and maximizing $|Q|^{-1} \sum_{i \in Q} p_i(y)$ over this set gives the joint interval and its width $W_Q^{\mathrm{aug}}(y)$. The query Q need not belong to the anchor family $\mathcal{I}$. Appendix I explains constraint normalization, and Appendix F gives an optional grouping rule based only on labeling-function outputs.

Simultaneous coverage and width. Coverage requires the global and local estimation errors to fall within their respective tolerances. We collect these requirements in the joint event $E := \{|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\} \cap \bigcap_{I \in \mathcal{I}} \{|\hat{\boldsymbol{\tau}}_I - \boldsymbol{\tau}_I| \preceq \boldsymbol{\lambda}_I\}$, where E means that every constraint is valid at the true labels. The budget δ_0 protects all global coordinates jointly, and δ_I protects all local coordinates for anchor I. The following proposition extends Proposition 1 by applying its truth-inclusion argument to the enlarged collection of constraints.

Proposition 6 (Validity). *On E, the true point-mass assignment $p_i^\circ(y) = \mathbb{1}\{y_i = y\}$ belongs to $\mathcal{U}^{\mathrm{aug}}$. Consequently, the augmented intervals cover $p_Q^\star(y)$ for all non-empty queries $Q \subseteq \{1, \dots, n\}$ and all $y \in \mathcal{Y}$. If the global event has failure probability at most δ_0 and each anchor event at most δ_I, then $\mathbb{P}(E) \geq 1 - \delta_0 - \sum_{I \in \mathcal{I}} \delta_I$.*

This union bound allows overlapping anchors and dependent estimation errors. It protects every queried group, including groups outside $\mathcal{I}$. The coverage and monotonicity proofs are in Appendix B. Coverage establishes that the truth is retained, while set inclusion determines how much the intervals can narrow.

Proposition 7 (Monotonicity at fixed tolerances). *For a non-empty $\mathcal{U}^{\mathrm{aug}}$ at the same $(\boldsymbol{\lambda}, \{\boldsymbol{\lambda}_I\})$, we have $\mathcal{U}^{\mathrm{aug}} \subseteq \mathcal{U}$ and $W_Q^{\mathrm{aug}}(y) \leq W_Q(y)$ for every non-empty query Q and label y. Under the prevalence anchor, additionally $W_I^{\mathrm{aug}}(1) \leq \min\{W_I(1), 2\lambda_I\}$ for every $I \in \mathcal{I}$.*

The $2\lambda_I$ bound applies to anchored groups. The inclusion comparison holds with the original global tolerances fixed. At a fixed total failure budget, allocating confidence to anchors increases the global tolerances, so the resulting empirical comparisons need not narrow every interval.

Precision from local labels. The labels that fall inside a group determine how tightly its anchor can constrain the target proportion. For the uniform draws J introduced above, let $\ell_I = \sum_{r=1}^{\ell} \mathbb{1}\{J_r \in I\}$ count group hits with multiplicity. For $\ell_I \geq 1$, let $\hat{\tau}_I$ be their positive-label fraction. The count $\ell_I \hat{\tau}_I$ is conditionally binomial with parameters ℓ_I and $p_I^\star(1)$. For the text construction, let $[L_I, H_I]$ be the Clopper–Pearson (CP) interval (Clopper & Pearson, 1934) at level $1 - \delta_I$ and take $\lambda_I = \max\{\hat{\tau}_I - L_I, H_I - \hat{\tau}_I\}$. The clipped symmetric anchor contains the raw CP interval. At zero hits we set $(\hat{\tau}_I, \lambda_I) = (1/2, 1/2)$, giving $[0, 1]$. The image construction uses $[L_I, H_I]$ directly, represented in (3) by its midpoint and half-width, with the same full-range zero-hit branch. The Clopper–Pearson radius satisfies the empirical Bernstein form below (Maurer & Pontil, 2009), with its endpoint bound proved in Appendix I.3. On average ℓ/ρ_I labeled draws reach I. The event $\ell_I \geq \ell/(2\rho_I)$ below controls precision through the available count; coverage is separately controlled by Proposition 6.

Theorem 8 (Group-size exponent). *For an anchored group I with a non-empty augmented set, suppose $\lambda_I \leq A_{\delta_I} \sqrt{\hat{p}(1-\hat{p})/\ell_I} + B_{\delta_I}/\ell_I$ when $\ell_I \geq 1$. Here $\hat{p} = \hat{\tau}_I$ is the empirical positive-label proportion. The constants are $A_\delta = \sqrt{2\log(2/\delta)}$ and $B_\delta = 3\log(2/\delta)$. On $\ell_I \geq \ell/(2\rho_I)$, an event of probability at least $1 - e^{-\ell/(8\rho_I)}$,*

$$W_I^{\mathrm{aug}}(1) \leq 2\lambda_I \leq 2A_{\delta_I} \sqrt{\frac{2\hat{p}(1-\hat{p})\rho_I}{\ell}} + \frac{4B_{\delta_I}\rho_I}{\ell}, \tag{4}$$

where the two terms on the right are the variance-dependent contribution and the finite-sample correction.

To compare this bound with global precision, suppose the global center $\hat{\boldsymbol{\tau}}$ is realizable and $\lambda_k = \Theta(\ell^{-1/2})$. Then Theorem 3 gives $W_I(1) \geq \min\{1, \Theta(\rho_I \ell^{-1/2})\}$, with full-range intervals whenever

$|I| \le n\lambda_{\min}$. Assume $1 \ll \rho_I \ll \sqrt{\ell}$ and the second term is lower order. In this non-saturated range, $W_I^{\mathrm{aug}}/W_I = O(\rho_I^{-1/2}) = O(\sqrt{|I|/n})$. Here the ratio refers to the widths for label 1 on the sufficient-count event.

For interior proportions with a dominant variance term, the anchored bound decreases as the inverse square root of the local count. Near the boundaries or with few hits, the finite-sample correction can dominate. Appendix I develops the binomial construction and the distinction between interior and boundary proportions, with a finite-population lower bound for interior proportions in Appendix C. We next extend the anchored construction to groups learned from labels.

5 Adaptive selection and independent estimation

Choosing a query from an already valid uncertainty set preserves its simultaneous coverage. Learning new anchors from the labels used to estimate them changes the confidence event. We now separate these two uses of labels. Independent estimation preserves the sampling relation after selection, giving coverage for the learned groups without restricting selector complexity.

Independent estimation after selection. To preserve the binomial relation for a learned group, we estimate its anchor from labels independent of the selector (Wasserman & Roeder, 2009). Let h be a predictor fitted to labeled examples that assigns a positive-class score $h(\Lambda_i) \in [0,1]$. For a threshold $t \in [0,1]$, define $I^+_{h,t} = \{i : h(\Lambda_i) \ge t\}$ and $I^-_{h,t} = \{i : h(\Lambda_i) \le 1-t\}$. The negative-class score is $1-h$, so these groups select large scores for either class and may overlap. If the fitting labels also estimate the anchors, a selector can favor groups that appear unusually pure on those observed labels (Berk et al., 2013). We partition the draw indices into $S \dot\cup E = [\ell]$ before reading the labels, where $[\ell] = \{1, \dots, \ell\}$. The selection sample is $\mathcal{D}_S = \{(\Lambda_{J_r}, y_{J_r}) : r \in S\}$ and the estimation sample $\mathcal{D}_E$ is defined analogously. The draw indices are disjoint and independent, although the same fixed pool instance can occur in both samples. We fit the selector and construct $\mathcal{G} = \mathcal{G}(\Lambda, \mathcal{D}_S)$ using only the selection sample and the pool's labeling-function outputs. Here Λ abbreviates the full collection $(\Lambda_1, \dots, \Lambda_n)$. We use $\mathcal{D}_E$ to estimate the anchors for $\mathcal{G}$. For each selected group, let $\ell_I^E = \sum_{r \in E} \mathbb{1}\{J_r \in I\}$ and let $\hat\tau_I^E$ be its empirical positive-label proportion when $\ell_I^E \ge 1$. We write $p_I^\star$ for $p_I^\star(1)$ in the following binomial relation. Let δ_I be the local failure budget and δ_{loc} the total budget for all anchors. We obtain $C_I^{\mathrm{split}}(y)$ by optimizing the group target over (3) with family $\mathcal{G}$ and Clopper–Pearson anchors estimated from the draws of $\mathcal{D}_E$ that fall in I.

Theorem 9 (Independent estimation after selection)**.** *Conditionally on $\mathcal{D}_S$, the family $\mathcal{G}$ is fixed. For any $I \in \mathcal{G}$ with $\ell_I^E \ge 1$, its positive-label count satisfies $\ell_I^E \hat\tau_I^E \mid (\mathcal{D}_S, \ell_I^E) \sim \mathrm{Bin}(\ell_I^E, p_I^\star)$, where conditioning fixes the selected group and its number of estimation draws. Local budgets δ_I are chosen before reading $\mathcal{D}_E$ and may depend on $\mathcal{D}_S$ provided $\sum_{I \in \mathcal{G}} \delta_I \le \delta_{\mathrm{loc}}$ almost surely. If the global moment event has failure probability at most δ_0, then $\mathbb{P}\{\forall I \in \mathcal{G}, \forall y \in \mathcal{Y} :\ p_I^\star(y) \in C_I^{\mathrm{split}}(y)\} \ge 1 - \delta_0 - \delta_{\mathrm{loc}}$. Groups with zero estimation draws receive a full-range local constraint.*

Here $C_I^{\mathrm{split}}(y)$ denotes the two joint LP endpoints, not the local confidence interval alone. A zero-hit anchor adds no information, while other constraints may still narrow its joint interval. The conditional coverage proof is in Appendix D. Once selection is fixed, the estimation draws that reach a group are independent uniform draws from that group. The guarantee therefore accommodates arbitrary selector complexity and a selection-dependent number of groups through the budget allocation.

Constructing and estimating learned groups. We fit a selector on $\mathcal{D}_S$, threshold its scores and deduplicate the resulting groups without using $\mathcal{D}_E$. The pattern selector h_1 returns a Laplace-smoothed positive rate for each output pattern. The naive Bayes selector h_2 pools class-conditional parameters across patterns. The logistic selector h_3 uses L_2 regularization on one-hot labeling-function outputs. A weight w mixes the fitted scores with the selection-sample prevalence. Appendix J specifies the score mixture, threshold grid and minimum group size. These selectors let us compare pattern-specific fitting with pooled estimation on the same inputs. We estimate each selected group's prevalence from $\mathcal{D}_E$ and add its confidence row to (1). We use all J to estimate fixed global features and allocate $\delta_0 = \delta/2$ across their coordinates. The remaining budget is divided equally among distinct anchors selected before estimation: the text family includes base and learned groups, whereas the image family contains the learned groups. This is a Bonferroni allocation. With a scalar anchors,

Table 1: Experimental batches. A repetition redraws the evaluation pool and labeled positions. Budgets count uniform draws with replacement, including repeated positions. Fixed-family width observations are group–repetition pairs, whereas coverage is assessed once per repetition and selector setting.

Batch	Corpora	Pool n	Draws ℓ	Repetitions	Reported interval
Fixed text	5	1600	100, 200, 400, 800	40	Joint LP
Size curves	5	1600	400	20	Joint LP
Coverage extension	5	1600	400	100	Joint LP
Learned text	8	1600	400	80	Symmetric local CP
Learned images	2	4000	200, 400, 800	60	Raw local CP

there are $2d + 2a$ inequalities in addition to probability bounds and two LP solves per query. The implementation aggregates identical output patterns; Appendix J specifies its variables and full-range fallback when an endpoint solve fails.

The learned-group results in Section 6 report *local-only intervals*: clipped symmetric CP anchors for text and raw CP intervals for images. Appendix J separately evaluates joint LP intervals. For precision, we compare split estimation with an all-draw width reference on the same groups; that reference reuses selection labels.

6 Experiments

We test two consequences of the analysis: whether local information improves precision on fixed groups, and whether independent estimation protects groups learned from labels. The comparisons separate interval width from simultaneous coverage.

6.1 Experimental setup

Fixed-group experiments use WRENCH labeling functions (Zhang et al., 2021) on IMDB (Maas et al., 2011), SMS, Yelp, TREC and AGNews (Zhang et al., 2015; Li & Roth, 2002). Anchors comprise individual output patterns, vote-ranked unions of patterns and vote-proportion threshold groups, all constructed without true labels. Learned-group experiments add SemEval (Hendrickx et al., 2010), Spouse and YouTube, using the selectors of Section 5. Image experiments test whether the selection effect persists with a different source of weak supervision: human annotation streams from CIFAR-10N and CIFAR-100N (Wei et al., 2022; Krizhevsky, 2009), combined with feature-based weak predictors. Table 1 gives pool sizes, labeled-draw budgets and repetitions; Appendix G specifies binary targets, feature construction and source splits.

For fixed groups, let G_0 denote the original global features, G_1 add pool prevalence, and A add local anchors to G_1. All use total failure budget $\delta = 0.1$. The global-only construction assigns the entire budget to global coordinates, whereas A divides it equally between the global event and all anchors. For learned groups, *same* estimates anchors on the selection half, *split* on the independent half, and *full* on all draws, with the selected groups held fixed. We report local-only intervals here: symmetric CP anchors for text and raw CP intervals for images. The full construction is a width reference that reuses selection labels; Table A5 separately reports joint LP results.

Metrics. A unit is one corpus–repetition under a fixed selector and budget, and fails simultaneous coverage if any queried proportion is uncovered. Fixed-group *tightening* is the median paired W_{G_1}/W_A, so values above one favor anchoring. Learned-group *inflation* is the median across units of median split width divided by median full width on the same groups, so smaller values mean a lower precision cost. Group–repetition pairs describe widths and are not independent coverage repetitions.

6.2 Local constraints improve subgroup precision

At $\ell = 400$, anchoring improves the median width on every corpus, with tightening factors of 1.21–1.69 (Table 2). The four-budget scan in Figure A3 shows larger median gains as more draws become available. Adding pool prevalence alone leaves its median factor near one, indicating that the

Table 2: **Anchoring improves median precision under a common total budget.** Coverage counts use 40 repetitions at $\ell = 400$: original global features (G_0) and joint anchored intervals (A). Tightening factors are medians of paired W_{G_1}/W_A, where G_1 adds pool prevalence to G_0 and matches the anchored global features. Each construction uses total $\delta = 0.1$.

Corpus	G_0 covered	A covered	Factor, $\ell = 100$	Factor, $\ell = 400$
IMDB	40	39	1.08	1.55
SMS	39	40	1.00	1.24
Yelp	40	40	1.00	1.21
TREC	39	39	1.13	1.68
AGNews	40	40	1.11	1.69

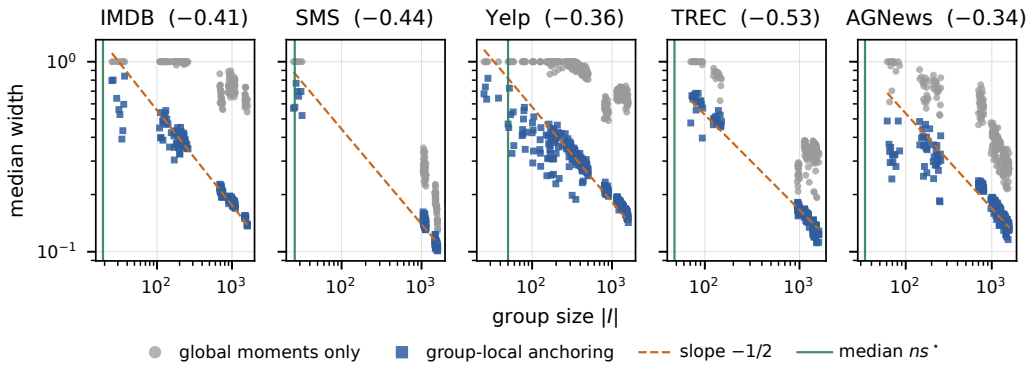


Figure 1: **Larger groups supply more local information.** Median widths at each observed group size, 20 repetitions, $\ell = 400$. Gray circles: original global features G_0; blue squares: anchored A. Green lines are median sufficient saturation sizes for G_0 from the separate 40-repetition batch under the same pool and feature configuration, used as descriptive references. Dashed lines show slope $-1/2$; titles report fitted joint-width slopes.

principal median improvement comes from local anchors (Table A11). At fixed global tolerances, no measured pair widens; with the common total budget, some do, as the global confidence allocation changes (Table A14).

Coverage distinguishes useful precision from overly narrow intervals. The jointly calibrated construction retains high measured coverage, including the saved 100-repetition extension (Table A10). Coordinatewise Wilson and standard-error tolerances have lower simultaneous coverage in the calibration ablation (Table A9); the joint guarantee itself follows from Proposition 6. Finite-pool adaptations of PPI (Angelopoulos et al., 2023a) and StratPPI (Fisch et al., 2024) often return zero-width intervals with zero estimated residual variance. Variance floors improve their observed coverage, while retaining a different normal-approximation calibration (Tables A3 and A4). These results locate the difficulty in the tested sparse-group variance estimates and motivate exact binomial calibration.

6.3 Group size determines the supply of local labels

Figure 1 shows decreasing anchored widths as group size increases. At the descriptive cut $|I| = 32$, median tightening rises from 1.03–1.12 below the cut to 1.77–2.41 above it (Table A12). The green lines address a separate question: Theorem 3 supplies a sufficient size for full-range global intervals, not an exact onset of anchoring gains. All measured groups below their own global threshold have full-range intervals (Table A14). Tolerance slopes lie near $-1/2$, whereas joint-width slopes also reflect clipping and other constraints (Table A13). The local-only width binds in most evaluated pairs, linking the observed precision primarily to how many labeled draws reach each group.

6.4 Independent estimation protects learned text groups

To isolate label reuse, we estimate each selected family's anchors from either the selection half or an equally sized independent half. For the unshrunk pattern selector, failures fall from 156 to 3 among 640 matched units (Table 3). The domain breakdown concentrates same-sample failures

Table 3: **Independent estimation reduces learned-text-group failures.** Local-only symmetric CP intervals, 640 units per row (eight corpora $\times$ 80 repetitions), $\ell = 400$. Same and split count units with any uncovered group. Inflation is the median across units of median split width divided by median full-draw width on identical groups. The full reference reuses selection labels.

Selector	Same failures	Split failures	Split/full inflation
Pattern, $w = 1$	156	3	1.62
Pattern, $w = 0.5$	113	2	1.56
Naive Bayes	4	2	1.54
Logistic	4	1	1.57

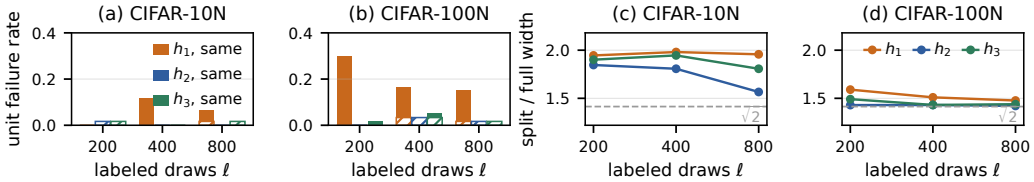


Figure 2: **Label reuse and estimation cost across image budgets.** Local-only raw CP intervals, 60 repetitions per setting. Panels (a,b): unit failure rates for same (solid) and split (hatched) estimation. Panels (c,d): median split/full width inflation. Selectors h_1, h_2, h_3 are pattern, naive Bayes and logistic. The dashed $\sqrt{2}$ line is an asymptotic half-sample reference.

in AGNews, Yelp and YouTube (Table A6). This pattern is consistent with fitting and evaluating separate positive rates for individual output patterns on the same observations. Pooling parameters yields fewer same-sample failures for the two other selectors tested.

Mixing the pattern score equally with selection-sample prevalence ($w = 0.5$) tests whether score shrinkage removes the dependence. It reduces same-sample failures but leaves many, whereas independent estimation restores the conditional binomial construction of Theorem 9. Across the tested selectors, split/full inflation is similar and corresponds to roughly 54%–62% wider intervals. This is the observed precision cost of using half the draws for estimation; group membership, local counts and confidence allocation also affect it (Tables A7 and A8).

6.5 Coverage and Width on Image Groups

Figure 2 varies the number of labeled draws to test the selection effect beyond text. On CIFAR-100N, the pattern selector has more same-sample failures than either pooled selector at all three budgets, and splitting reduces them. CIFAR-10N shows a smaller difference: the pattern selector has no same-sample failures at 200 draws, with counts of 7 and 4 at the larger budgets (Table A2). These counts describe the observed differences, not a uniform claim of excess failure across budgets. Split/full inflation ranges from 1.42 to 1.98, reflecting local sample counts and discrete interval construction beyond the leading $\sqrt{2}$ scale.

7 Conclusion

Global confidence constraints can give simultaneous coverage while leaving small-group proportions unresolved. Our width bound identifies a computable sufficient condition for full-range intervals, motivating constraints estimated from the labels reaching each group. Joint calibration preserves coverage, and splitting labeled draws extends this construction to learned groups. The experiments show median precision gains where local information is available and reduced failures from label reuse, with a measurable width cost from reserving an independent estimation sample. These results connect subgroup resolution to local label availability and the way those labels enter the procedure.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Verónica Álvarez, Santiago Mazuelas, Steven An, and Sanjoy Dasgupta. Reliable programmatic weak supervision with confidence intervals for label probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. arXiv:2508.03896.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. URL https://arxiv.org/abs/2311.01453.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

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.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision. In *Advances in Neural Information Processing Systems*, 2021.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973. doi: 10.1214/aos/1193342380.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John Duchi. Predictive inference with weak supervision. *arXiv preprint arXiv:2201.08315*, 2022. arXiv:2201.08315.

Yiqun T Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. *arXiv preprint arXiv:2601.05420*, 2026. URL https://arxiv.org/abs/2601.05420.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11493–11514. PMLR, 2025. URL https://proceedings.mlr.press/v267/csillag25a.html.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3): 1902–1933, 2004.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. *arXiv preprint arXiv:2601.20913*, 2026. URL https://arxiv.org/abs/2601.20913.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111489–111514, 2024. doi: 10.52202/079017-3541.

Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. arXiv:2002.11955.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. arXiv:1512.03385.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. *arXiv preprint arXiv:2511.21140*, 2025. URL https://arxiv.org/abs/2511.21140.

Puheng Li, Tijana Zrnic, and Emmanuel J. Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, 2025.

Sida Li and Nikolaos Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34836–34875. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25ak.html.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135704–135740, 2024. doi: 10.52202/079017-4309.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003. ISBN 978-0-387-00454-9.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. arXiv:0907.3740.

Alessio Mazzetto, Dylan Sam, Andrew Park, Eli Upfal, and Stephen Bach. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Alessio Mazzetto, Reza Esfandiarpoor, Akash Singirikonda, Eli Upfal, and Stephen H. Bach. An adaptive method for weak supervision with drifting data. *arXiv preprint arXiv:2306.01658*, 2023. arXiv:2306.01658.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934.

Sangwoo Park, Matteo Zecchin, and Osvaldo Simeone. Adaptive prediction-powered autoeval with reliability and efficiency guarantees. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-5303. URL https://doi.org/10.52202/085713-5303.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, 2016. arXiv:1605.07723.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment*, volume 11, pp. 269–282, 2017. arXiv:1711.10160.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. arXiv:1810.02840.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. ISBN 978-0-387-00152-4.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of Statistics*, 37 (5A):2178, 2009.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. arXiv:2110.12088.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A comprehensive benchmark for weak supervision. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. arXiv:2109.11377.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022. arXiv:2202.05433.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A PROOF OF THEOREM 3

This appendix proves the global-width lower bound. We restate Theorem 3 before its perturbation argument.

> **Theorem 3 (restated).**
> Let $\mathcal{U}$ be non-empty and $\Phi_k(\Lambda_i, y) \in [0,1]$ for all i and k and $y \in \mathcal{Y}$. Then every non-empty $I \subseteq \{1,\dots,n\}$ and every $y \in \{0,1\}$ satisfy
> $$W_I(y) \geq \min\{1,\ \rho_I s^\star\}, \qquad \rho_I = n/|I|,$$
> where $s^\star$ is the center slack of Definition 2. If $\hat{\boldsymbol{\tau}}$ is realizable, meaning $g(p) = \hat{\boldsymbol{\tau}}$ for some $p \in \Delta(\mathcal{Y})^n$, then $s^\star = \min_k \lambda_k$.

Take $y = 1$ without loss of generality: for $T = 2$ we have $p_i(0) = 1 - p_i(1)$, so $[\underline{p}_I(0), \overline{p}_I(0)] = [1 - \overline{p}_I(1), 1 - \underline{p}_I(1)]$ and the two widths coincide.

Let $p^c \in \mathcal{U}$ attain $s^\star$ and put $q := \frac{1}{|I|}\sum_{i\in I} p_i^c(1)$. For $t \in [0,1]$ define an upward perturbation supported on I by $p_i^{t+}(1) := p_i^c(1) + t(1 - p_i^c(1))$ for $i \in I$ and $p_i^{t+} := p_i^c$ otherwise, which stays in the simplex pointwise. Its moment is $g(p^{t+}) = g(p^c) + t\,a$ with $a = \frac{1}{n}\sum_{i\in I}(1 - p_i^c(1))\Delta_i$, and since $1 - p_i^c(1) \in [0,1]$ and $|\Delta_{i,k}| \leq 1$,

$$|a_k| \ \leq\ \frac{1}{n}\sum_{i\in I} 1 \ =\ \frac{|I|}{n} \ =\ \rho_I^{-1} \qquad \text{for every } k. \tag{5}$$

Coordinatewise we obtain $|g_k(p^{t+}) - \hat{\tau}_k| \leq (\lambda_k - s_k(p^c)) + t|a_k| \leq \lambda_k - s^\star + t\rho_I^{-1}$, using $s_k(p^c) \geq s^\star$. The right-hand side is at most λ_k whenever $t \leq \rho_I s^\star$. Writing $t_\star := \min\{1, \rho_I s^\star\}$ we get $p^{t_\star +} \in \mathcal{U}$ and hence $\overline{p}_I(1) \geq q + t_\star(1 - q)$. The downward family $p_i^{t-}(1) := (1-t)p_i^c(1)$ for $i \in I$ shifts the moment by $t\,b$ with $b = -\frac{1}{n}\sum_{i\in I} p_i^c(1)\Delta_i$, which obeys the same bound (5), so $p^{t_\star -} \in \mathcal{U}$ and $\underline{p}_I(1) \leq q - t_\star q$. Subtracting gives $W_I(1) \geq t_\star(1-q) + t_\star q = t_\star$.

For the realizable case, any p with $g(p) = \hat{\boldsymbol{\tau}}$ lies in $\mathcal{U}$ and has $s_k(p) = \lambda_k$, so $s^\star \geq \min_k \lambda_k$. The reverse holds because $s_k(p) \leq \lambda_k$ for every feasible p. □

A.1 PERTURBATION INTERPRETATION AND LABEL-BUDGET SCALING

Changing instance i from label 0 to label 1 changes its feature vector by $\Delta_i := \boldsymbol{\Phi}(\Lambda_i, 1) - \boldsymbol{\Phi}(\Lambda_i, 0)$. These binary feature differences control changes in the pool moment. The feature range gives $\|\Delta_i\|_\infty \leq 1$.

The proof above moves group probabilities toward the two binary extremes from a point attaining $s^\star$. A common interpolation step changes each global moment by at most that step divided by ρ_I. The upward and downward changes in the group average sum to the step. Thus the admissible step gives the interval width in (2).

To express this result in terms of the label budget, let $\lambda_{\min} := \min_k \lambda_k$. Standard-error and Wilson tolerances (Wilson, 1927) have square-root sample-size scaling when the estimated moments have non-zero limiting variance. The following corollary substitutes that scaling into the realizable-center case.

Corollary 10 (Amplification exponent)**.** *Suppose $\hat{\boldsymbol{\tau}}$ is realizable and the tolerance schedule satisfies $\lambda_k = \Theta(\ell^{-1/2})$. Then $W_I(y) \geq \min\{1, \Theta(\rho_I \ell^{-1/2})\}$. In particular, every group with $|I| \leq n\lambda_{\min}$ has the full interval $[0,1]$.*

Corollary 10 follows directly from (2). It shows that the width lower bound grows linearly with ρ_I until saturation. Across the five corpora at $\ell = 400$, the realizable-center fraction ranges from 0 to 0.6 over 40 repetitions per corpus while nonrealizable cases have corpus-median gaps $\lambda_{\min} - s^\star$ from 0 to 0.0032 and maximum gap 0.018. Solving the zero-tolerance equality program directly gives realizable counts $18/0/24/2/0$ out of 40 in the order IMDB/SMS/Yelp/TREC/AGNews, and adding the pool-prevalence coordinate leaves every one of these counts unchanged. The empirical resolution analysis therefore uses the general slack form of Theorem 3 rather than the realizable special case.

Call coordinate k *unipolar* when $\Phi_k(\Lambda_i, y) = \mathbb{1}\{i \in I_k\}\mathbb{1}\{y = y_k\}$, where I_k is the firing group fixed by the labeling-function outputs and y_k is the voted label. A keyword rule that votes positive whenever its keyword appears gives such a coordinate. For a unipolar coordinate, the corresponding global row directly bounds the average probability of y_k within I_k. Its interval length is $2\rho_{I_k}\lambda_k$, giving $W_{I_k}(y_k) \le \min\{1, 2\rho_{I_k}\lambda_k\}$. Combined with Theorem 3, this matches the exponent of one before saturation when the center is realizable and $\lambda_k \asymp \lambda_{\min}$.

A.2 Center slack and its computation

The set in (1) constrains pool averages, whereas its endpoints optimize group averages. We quantify the resulting width by measuring the margin available inside the global constraints. The global-width analysis also identifies a common labeling-function form for which the group-size dependence has a matching upper bound.

To turn this sensitivity into a width bound, we need the distance a feasible moment can move before reaching a constraint. We measure that distance by the smallest remaining margin across coordinates.

The center slack is the largest margin available simultaneously in every coordinate. It equals zero when every feasible moment touches the tolerance-box boundary. Introducing a variable bounded above by each coordinate margin gives a linear program comparable in size to an endpoint program. Theorem 3 translates this available margin into unavoidable group-level variation.

B Proofs for Section 4

This appendix proves the coverage and width results for the anchored set in Section 4. Each statement is repeated before its proof.

B.1 Validity

Proposition 6 (restated).
On E, the true point-mass assignment $p_i^\circ(y) = \mathbb{1}\{y_i = y\}$ belongs to $\mathcal{U}^{\mathrm{aug}}$. Consequently, the augmented intervals cover $p_I^\star(y)$ for all non-empty $I \subseteq \{1, \dots, n\}$ and all $y \in \mathcal{Y}$. If the global event has failure probability at most δ_0 and each anchor event at most δ_I, then $\mathbb{P}(E) \ge 1 - \delta_0 - \sum_{I \in \mathcal{I}} \delta_I$.

The point mass p° satisfies $\frac{1}{|I|}\sum_{i \in I} \mathbb{E}_{p_i^\circ} \mathbf{\Phi}^{\mathrm{loc}}(\Lambda_i, \cdot) = \boldsymbol{\tau}_I$ by definition, so on E every anchor row in Definition 5 holds at p°. The global rows hold on E by the same substitution in (1). Hence $p^\circ \in \mathcal{U}^{\mathrm{aug}}$, and for any I and y the value $p_I^\star(y) = \frac{1}{|I|}\sum_{i \in I} p_i^\circ(y)$ lies between the minimum and the maximum of that linear functional over a set containing p°. The probability statement is a union bound. □

B.2 Monotonicity at fixed tolerances

Proposition 7 (restated).
For a non-empty $\mathcal{U}^{\mathrm{aug}}$ at the same $(\boldsymbol{\lambda}, \{\boldsymbol{\lambda}_I\})$, we have $\mathcal{U}^{\mathrm{aug}} \subseteq \mathcal{U}$ and $W_I^{\mathrm{aug}}(y) \le W_I(y)$ for every non-empty group I and label y. Under the prevalence anchor, additionally $W_I^{\mathrm{aug}}(1) \le \min\{W_I(1), 2\lambda_I\}$ for every $I \in \mathcal{I}$.

A maximum over a subset does not increase and a minimum does not decrease, which gives $\mathcal{U}^{\mathrm{aug}} \subseteq \mathcal{U}$ and the width comparison. For the second part, the anchor row confines the objective $\frac{1}{|I|}\sum_{i \in I} p_i(1)$ to an interval of length $2\lambda_I$, so the difference of its maximum and minimum over $\mathcal{U}^{\mathrm{aug}}$ is at most $2\lambda_I$. □

B.3 Group-size dependence

Theorem 8 (restated).
For an anchored group I with a non-empty augmented set, suppose $\lambda_I \le A_{\delta_I}\sqrt{\hat{p}(1-\hat{p})/\ell_I} + B_{\delta_I}/\ell_I$ when $\ell_I \ge 1$. Here $\hat{p} = \hat{\tau}_I$ is the empirical positive-label proportion. The constants are $A_\delta = \sqrt{2\log(2/\delta)}$ and $B_\delta = 3\log(2/\delta)$.

Then with probability at least $1 - e^{-\ell/(8\rho_I)}$,

$$W_I^{\mathrm{aug}}(1) \le 2\lambda_I \le 2A_{\delta_I}\sqrt{\frac{2\hat{p}(1-\hat{p})\rho_I}{\ell}} + \frac{4B_{\delta_I}\rho_I}{\ell},$$

where the two terms on the right are the variance-dependent contribution and the finite-sample correction. Suppose the global center $\hat{\boldsymbol{\tau}}$ is realizable and $\lambda_k = \Theta(\ell^{-1/2})$. Then Theorem 3 gives $W_I(1) \ge \min\{1, \Theta(\rho_I \ell^{-1/2})\}$, with full-range intervals whenever $|I| \le n\lambda_{\min}$. Assume $1 \ll \rho_I \ll \sqrt{\ell}$ and the second term is lower order. In this non-saturated range, $W_I^{\mathrm{aug}}/W_I = O(\rho_I^{-1/2}) = O(\sqrt{|I|/n})$. Here the ratio refers to the widths for label 1.

Proposition 7 gives $W_I^{\mathrm{aug}} \le 2\lambda_I$. Draws are i.i.d. uniform on $\{1, \dots, n\}$, so conditionally on falling in I a draw is uniform on I and draws remain independent. Given $\ell_I = \ell'$ the labels seen inside I are i.i.d. Bernoulli with parameter $p_I^\star(1)$. Hence $\ell_I \sim \mathrm{Bin}(\ell, \rho_I^{-1})$ and the multiplicative Chernoff bound gives $\mathbb{P}\{\ell_I < \ell/(2\rho_I)\} \le e^{-\ell/(8\rho_I)}$. On the complementary event substitute $\ell_I \ge \ell/(2\rho_I)$ into the assumed form of λ_I, which yields (4). Dividing $\sqrt{\rho_I/\ell}$ by $\rho_I/\sqrt{\ell}$ gives $\rho_I^{-1/2}$. □

B.4 Consequences of truth inclusion and set restriction

The union bound allows overlapping groups and dependent estimation errors. It also implies that an empty $\mathcal{U}^{\mathrm{aug}}$ certifies that the joint confidence event did not occur.

The monotonicity proof above uses the fact that restricting a feasible set raises minima and lowers maxima. For an anchored group, the target also lies within a row of length $2\lambda_I$.

Three reference frames measure how often anchoring widens an interval, and each uses the same $4{,}540$ anchored group-repetition pairs at $\ell = 400$ under Bonferroni–Clopper–Pearson tolerances. At fixed global tolerances 0 pairs widen, which is what Proposition 7 requires. Against the same global features under the divided failure budget, 214 pairs widen. Against the global features without the pool-prevalence coordinate, 8 pairs widen.

C A finite-population lower bound

This appendix supports the local-sample-size comparison in Section 4 and the finite-population scope discussed in Appendix L.

Consider a group of finite-population size $N = |I|$ observed through ℓ_I draws. A two-point argument in the style of Cam (1973) applies for any $0 < \gamma < 1 - 2\delta$ whose two comparison proportions both lie in the interior of $[0, 1]$. The width ω of any interval meeting the coverage requirement at such a proportion satisfies

$$\omega \;\ge\; \frac{\gamma}{\sqrt{2\ell_I}} \;-\; \sqrt{\frac{2}{N}\log\Big(\frac{2}{1-2\delta-\gamma}\Big)}. \tag{6}$$

The leading term is the unavoidable $\ell_I^{-1/2}$ scale and the second is the finite-population correction. The bound is non-vacuous only when N/ℓ_I exceeds a threshold $\Gamma(\delta)$. The numerical thresholds are 15.02 at $\delta = 0.01$ and 18.41 at $\delta = 0.05$. They are 24.34 at $\delta = 0.10$ and 47.62 at $\delta = 0.20$. Our text settings have $N/\ell_I \approx 4$, where (6) is vacuous. The comparison establishes the asymptotic $1/2$ scale. The upper and lower constants differ by a factor of 7.6 to 8.5, so their optimal values remain undetermined.

C.1 Comparison with the local estimation radius

To assess whether local labels can support a faster estimation scale, we use a finite-population two-point argument following Cam (1973). The argument compares two interior label proportions whose sampled observations are difficult to distinguish. Equation (6) gives the resulting lower bound with its finite-population correction. It identifies the $\ell_I^{-1/2}$ asymptotic scale for inference from local labels. For the allocations used in this comparison, the local-radius coefficient A_{δ_I} is 7.6 to 8.5 times the leading lower-bound coefficient. This comparison concerns the asymptotic rate, with the finite-population range discussed in Appendix L.

D Proof of Theorem 9

This appendix proves the conditional sampling and coverage statements in Section 5. We repeat Theorem 9 before conditioning on the selection sample.

> **Theorem 9 (restated).**
> Conditionally on $\mathcal{D}_S$, the family $\mathcal{G}$ is fixed. For any $I \in \mathcal{G}$ with $\ell_I^E \geq 1$, its positive-label count satisfies
>
> $$\ell_I^E \hat{\tau}_I^E \mid (\mathcal{D}_S, \ell_I^E) \sim \mathrm{Bin}\big(\ell_I^E,\, p_I^\star\big),$$
>
> where conditioning fixes the selected group and its number of estimation draws. Local budgets δ_I may depend on $\mathcal{D}_S$ provided $\sum_{I\in\mathcal{G}} \delta_I \leq \delta_{\mathrm{loc}}$ almost surely. If the global moment event has failure probability at most δ_0, then $\mathbb{P}\{\forall I \in \mathcal{G}, \forall y \in \mathcal{Y} :\ p_I^\star(y) \in C_I^{\mathrm{split}}(y)\} \geq 1 - \delta_0 - \delta_{\mathrm{loc}}$. Groups with zero estimation draws receive a full-range local constraint.

Condition on $\mathcal{D}_S$. The family $\mathcal{G} = \mathcal{G}(\Lambda, \mathcal{D}_S)$ is then a fixed collection of subsets of $\{1, \dots, n\}$, and so are the local budgets δ_I. The positions $\{J_r\}_{r\in E}$ are drawn independently of $\{J_r\}_{r\in S}$ and of the labels revealed at those positions, and each is uniform on $\{1, \dots, n\}$. Hence conditionally on $\mathcal{D}_S$ and on ℓ_I^E, the positions in E that land in I are uniform on I and mutually independent. The number of ones among them is $\mathrm{Bin}(\ell_I^E, p_I^\star)$ as stated in Theorem 9. A Clopper–Pearson interval at level $1 - \delta_I$ is therefore valid conditionally on $\mathcal{D}_S$, and $\ell_I^E = 0$ leaves the row vacuous. Let F_I denote failure of the local constraint for I and F_0 failure of the global moment event. Conditional on $\mathcal{D}_S$, the union bound gives $\mathbb{P}(\bigcup_{I\in\mathcal{G}} F_I \mid \mathcal{D}_S) \leq \sum_{I\in\mathcal{G}} \delta_I \leq \delta_{\mathrm{loc}}$. Taking expectations gives $\mathbb{P}(\bigcup_{I\in\mathcal{G}} F_I) \leq \delta_{\mathrm{loc}}$. The global guarantee is unconditional, with $\mathbb{P}(F_0) \leq \delta_0$. Consequently $\mathbb{P}(F_0 \cup \bigcup_{I\in\mathcal{G}} F_I) \leq \delta_0 + \delta_{\mathrm{loc}}$. On the complementary event the true assignment satisfies every row, which gives the stated simultaneous coverage. □

E Multiclass extension

This appendix gives the multiclass extensions of the global-width argument and their numerical comparisons.

For K classes the group target is $q_I^y(p) = \frac{1}{|I|}\sum_{i\in I} p_i(y)$ and the simplex constraint couples classes: raising one class inside a group must take mass from the others, so the symmetric two-point perturbation of Appendix A needs to be replaced. Three nested bounds result. The uniform bound reuses the uniform coordinate sensitivity and reads $W_I(y) \geq \min\{1, \rho_I s^\star\}$. The second replaces ρ_I^{-1} by a group- and class-specific sensitivity $C_I(y) \leq |I|/n$, giving $W_I(y) \geq \min\{1, s^\star/C_I(y)\}$. The third constructs explicit upward and downward witness directions with feasible step lengths $t_\pm$ and yields $W_I(y) \geq t_+(1-q) + t_- q$ at the center value $q = q_I^y(p^c)$. For the majority-vote feature map used in the anchor construction no coordinate abstains, so $C_I(y)$ attains $|I|/n$ and the second layer collapses onto the uniform bound. The witness layer still exploits directional asymmetry. On a grid of $7{,}497$ points at two label budgets there are no violations of any layer, and the median gain of the witness layer over the uniform layer is 1.260 at $\ell = 400$ and 1.000 at $\ell = 100$, where more intervals are already saturated. A further $29{,}231$ multiclass lower-bound checks on text corpora also record no violation. These numerical comparisons agree with the stated analytic bounds. The analytic argument establishes the guarantee, while the numerical comparisons assess its implementation.

F Predictor-assisted grouping as a constructive admissible family

This appendix specifies predicted-width adaptive grouping (PAG), an optional admissible family. For each distinct labeling-function pattern u, let q_u be the fraction of its nonabstaining votes that are positive, using $q_u = 1/2$ for an all-abstaining pattern. Order patterns by q_u, breaking ties by pattern index, and consider contiguous segments of this order. For a segment I, let $\bar{q}_I$ be the instance-weighted mean of q_u and $\hat{\ell}_I = \ell|I|/n$. For a candidate partition into k segments, allocate

Table A1: Predicted-width grouping versus equal-count boundaries, $\ell = 400$. Both use the same pattern order and number of groups, but define different targets. The width ratio is the median across repetitions of the adaptive median width divided by the equal-count median width. Coverage counts are simultaneous over each partition.

Corpus	Repetitions	Median groups	Adaptive/equal	Adaptive covered	Equal covered
IMDB	40	2.5	1.01	39	39
SMS	40	2	1.00	39	39
Yelp	40	4	0.99	40	40
TREC	40	2	0.98	38	38
AGNews	40	4	0.92	39	39

$\delta_I = \delta/(2k)$ and predict

$$\widehat{W}_I = \min\left\{1, 2\left[\sqrt{\frac{2\log(2/\delta_I)\bar{q}_I(1-\bar{q}_I)}{\hat{\ell}_I}} + \frac{3\log(2/\delta_I)}{\hat{\ell}_I}\right]\right\}.$$

For each $k = 1, \ldots, \min(24, U)$, where U is the number of patterns, dynamic programming minimizes $\sum_I |I|\widehat{W}_I/n$ subject to $\widehat{W}_I \leq 0.5$ for every segment. The implemented search stops at its first infeasible k and retains the last feasible partition, or the whole pool if even $k = 1$ fails. All inputs are functions of the labeling-function outputs and the prespecified sampling budget, so Definition 4 applies without fitting to true labels. The comparison rule uses the same order and number of segments, with boundaries nearest equal instance counts.

Table A1 compares the resulting partitions with equal-count boundaries. Ratios near one on most corpora show little median precision advantage from optimizing this proxy at the tested budget. The partitions generally contain different groups, so the comparison describes their resulting precision at a matched number of groups.

G Data sources and complete image results

Table 1 separates the experimental batches used in Section 6. Fixed text groups use five corpora; learned text groups add three more to test selection on a broader range of labeling-function outputs. Individual-pattern groups collect identical output vectors, pattern-prefix groups unite patterns ranked by weak-vote purity, and vote-threshold groups select large positive or negative vote proportions. These form the fixed anchor family after deduplication. Label-dependent predictor groups in the fixed-family study are queried without adding their anchors.

Text corpora. We use the following source splits and binary targets. Each positive rate describes the corresponding source pool.

AGNews. The test split has 12,000 instances and 9 labeling functions. The target is class 0 against the rest. Its positive rate is 0.2366.

IMDB. The training split has 20,000 instances and 5 labeling functions. The task is binary with positive rate 0.4991.

SMS. The training split has 4,571 instances and 10 labeling functions. The task is binary with positive rate 0.1362.

TREC. The training split has 5,033 instances and 10 labeling functions. The target is class 0 against the rest. Its positive rate is 0.2099.

Yelp. The training split has 30,400 instances and 8 labeling functions. The task is binary with positive rate 0.4967.

YouTube. The training, validation and test splits have 2,056 instances and 10 labeling functions. The task is binary with positive rate 0.5151.

Spouse. The validation and test splits have 5,512 instances and 9 labeling functions. The task is binary with positive rate 0.0740.

SemEval. The training, validation and test splits have 2,527 instances. We retain ten of 164 labeling functions. The target is class 0 against the rest with positive rate 0.1591.

Table A2: Local-only raw CP intervals for all image settings, 60 units per row. Same and split are unit failure counts. Inflation is median split/full width inflation on identical groups. The final column is the median number of selected groups per unit.

Corpus	ℓ	Selector	Same	Split	Inflation	Groups
CIFAR-10N	200	Pattern	0	0	1.95	8
CIFAR-10N	200	Naive Bayes	0	1	1.85	6
CIFAR-10N	200	Logistic	0	1	1.90	16
CIFAR-10N	400	Pattern	7	0	1.98	9
CIFAR-10N	400	Naive Bayes	0	0	1.81	5
CIFAR-10N	400	Logistic	0	0	1.95	14
CIFAR-10N	800	Pattern	4	1	1.96	11
CIFAR-10N	800	Naive Bayes	0	0	1.57	5
CIFAR-10N	800	Logistic	0	1	1.81	14
CIFAR-100N	200	Pattern	18	0	1.59	12
CIFAR-100N	200	Naive Bayes	0	0	1.43	13
CIFAR-100N	200	Logistic	1	0	1.49	14
CIFAR-100N	400	Pattern	10	2	1.51	14
CIFAR-100N	400	Naive Bayes	1	2	1.43	13
CIFAR-100N	400	Logistic	3	2	1.43	15
CIFAR-100N	800	Pattern	9	1	1.48	15
CIFAR-100N	800	Naive Bayes	1	1	1.42	13
CIFAR-100N	800	Logistic	1	1	1.44	14.5

The Spouse training split is excluded because its labels are unavailable or uniformly negative. If a corpus has more than 10 labeling functions, we retain the 10 with the highest unlabeled coverage. This selection uses no true labels.

Image corpora. Each source training set of $50{,}000$ images is split by fixed index. We use $30{,}000$ images only to fit the weak feature predictors and the remaining $20{,}000$ as the sampling pool. Features come from ResNet-18.

For CIFAR-10N the fixed binary task is animals against vehicles. Original classes $0, 1, 8, 9$ are negative and the pool positive rate is 0.5997. The labeling functions use five human annotation streams: aggregate and worst plus three independent rounds. Their five-way agreement indicator and two ridge predictors give 8 coordinates in total. The pool coverages are $(1, 1, 1, 1, 1, 0.88165, 0.81855, 0.81305)$.

For CIFAR-100N the task is fine-label parity with pool positive rate 0.5014. The two human streams provide fine and coarse labels. Their agreement indicator and two ridge predictors give 5 coordinates. The pool coverages are $(1, 1, 0.58585, 0.71285, 0.52635)$. Clean labels enter the experiments only through the sampled draws, which estimate the global moments and the anchors and fit the selectors. They also define the true group proportions and the coverage events, and they never enter a labeling function.

Allocation and thresholds. Within a unit, half of δ is reserved for the global moment event and half is divided equally among the G distinct calibration groups in that unit, so $\delta_I = (\delta/2)/G$ for $G > 0$. Text calibration includes the base and learned families; image calibration includes only learned groups. When $G = 0$, no local constraints are added and the allocated global budget remains $\delta/2$. Thresholds run from 0.90 to 0.50 in steps of 0.05. Groups below 5 instances are dropped and duplicates by membership are removed.

The coverage counts and width ratios for every image setting are listed in Table A2. Section 6.5 discusses their dependence on selector choice and label budget.

H GLOBAL CONFIDENCE CONSTRAINTS

This appendix develops the joint-calibration example and inference-method comparison for global constraints.

H.1 JOINT CALIBRATION AND INFERENCE TARGETS

The confidence level in Proposition 1 applies jointly to all d coordinates. For example, independent coordinate events with individual coverage $1 - \delta$ have joint coverage $(1 - \delta)^d$. At $d = 10$ and $\delta = 0.1$, this is 0.35 rather than the desired $1 - \delta$. This calculation illustrates why we allocate the confidence budget across coordinates. Section 6 compares joint calibration with separate per-coordinate tolerances.

Relation to existing inference methods. Label models combine conflicting labeling-function votes into per-instance estimates (Ratner et al., 2016; 2017; 2019; Fu et al., 2020). The survey of Zhang et al. (2022) describes this design space. We instead study the range of group proportions compatible with confidence constraints. This is a partial identification perspective (Manski, 2003). Predictive inference under weak supervision builds label sets for a future exchangeable test point (Cauchois et al., 2022). Its marginal prediction guarantee complements our simultaneous intervals for proportions in a fixed finite pool. In addition, Mazzetto et al. (2023) study estimation error for labeling-function parameters under drift. Our width analysis concerns how tolerances on pool statistics translate into precision for a particular group. We build directly on the uncertainty set of Álvarez et al. (2025) and add constraints at group granularity.

H.2 ADDITIONAL CONTEXT FOR WEAK SUPERVISION AND INFERENCE

To combine these functions, data programming uses a label model to estimate the relationship between labeling-function outputs and unknown labels (Ratner et al., 2016). Building on this approach, Snorkel learns labeling-function accuracies and dependencies to combine conflicting votes into probabilistic training labels (Ratner et al., 2017). Among subsequent label models, MeTaL jointly models weak supervision across related tasks to estimate their latent labels (Ratner et al., 2019). Alternatively, FlyingSquid estimates label-model parameters from agreements among triplets of labeling functions to accelerate probabilistic labeling (Fu et al., 2020). These models provide per-instance probabilities whose interpretation depends on their assumptions about the labeling functions.

Separately, prediction-powered inference (PPI) uses a labeled sample to estimate prediction errors and correct estimates obtained from a larger prediction sample (Angelopoulos et al., 2023a). Its stratified variant StratPPI adapts this correction and sample allocation across subpopulations to improve statistical efficiency (Fisch et al., 2024). Our finite-pool adaptations use fixed majority-vote predictions $f_i \in \{0, 1/2, 1\}$, with ties assigned $1/2$. For group I, PPI estimates $|I|^{-1} \sum_{i \in I} f_i$ plus the mean residual $y - f$ among labeled hits. Its normal half-width is $z_{1-\alpha_I/2} s_I / \sqrt{\ell_I}$, where s_I^2 is sample residual variance and $\alpha_I = 0.1/G$ for G queried groups. With fewer than two hits it returns $[0, 1]$. StratPPI partitions weak-vote scores into $[0, .25), [.25, .5), [.5, .75), [.75, 1]$, allocating the same total of 400 draws equally across nonempty strata. It weights group–stratum residual corrections by their finite-pool shares and sums variances with squared weights. An intersection with fewer than two hits retains its full probability-range contribution. The predictor means are known over the fixed pool, and the prediction weight is one.

Tables A3 and A4 show coverage alongside widths and zero-width frequencies. At zero estimated residual variance, the normal radius vanishes even when the corrected estimate misses the target. Replacing s_I^2 by $\max\{s_I^2, c/\ell_I\}$ improves observed coverage and removes these zero-width intervals for the positive floors tested. This sensitivity analysis changes the variance estimate and does not establish finite-sample simultaneous validity for the normal intervals.

Weak-to-strong learning examines how a stronger model can improve upon the weak model supplying its training labels (Burns et al., 2024). This concerns predictive performance through weak training labels. Our interval analysis instead assesses uncertainty in selected groups.

False discovery rate control bounds the expected fraction of false rejections among the decisions rejected by the procedure (Benjamini & Hochberg, 1995). This error criterion differs from simultaneous coverage of all queried proportions. Sequential allocation methods also use variance information (Audibert et al., 2009). Self-normalized processes relate fluctuations to their observed scale (de la Peña et al., 2004), while betting constructions yield confidence sequences that remain valid over time (Waudby-Smith & Ramdas, 2024).

Table A3: Finite-pool PPI adaptation and variance-floor sensitivity, $\ell = 400$ and 40 repetitions per corpus. The floor replaces sample residual variance by $\max\{s^2, c/\ell_I\}$. Coverage counts are simultaneous, widths are group-pair medians and interquartile ranges, and zero width is a percentage of the listed group pairs. These normal intervals are an empirical comparator without the finite-sample guarantee of the anchored CP construction.

Corpus	c	Covered	Width	Width IQR	Zero (%)	Pairs
IMDB	0	1	0.33	[0.18, 0.72]	13.1	1033
IMDB	0.25	26	0.38	[0.22, 0.72]	0.0	1033
IMDB	1	33	0.42	[0.23, 0.90]	0.0	1033
SMS	0	1	0.00	[0.00, 0.20]	52.2	728
SMS	0.25	37	0.35	[0.13, 0.70]	0.0	728
SMS	1	40	0.57	[0.17, 1.00]	0.0	728
Yelp	0	0	0.29	[0.00, 0.98]	26.5	1095
Yelp	0.25	26	0.48	[0.25, 0.98]	0.0	1095
Yelp	1	35	0.65	[0.29, 1.00]	0.0	1095
TREC	0	3	0.15	[0.00, 0.48]	39.2	742
TREC	0.25	30	0.19	[0.10, 0.50]	0.0	742
TREC	1	37	0.28	[0.14, 0.56]	0.0	742
AGNews	0	0	0.14	[0.00, 0.31]	39.9	942
AGNews	0.25	31	0.24	[0.13, 0.55]	0.0	942
AGNews	1	39	0.29	[0.17, 0.96]	0.0	942

Table A4: Finite-pool StratPPI adaptation and variance-floor sensitivity, $\ell = 400$ and 40 repetitions per corpus. The floor replaces sample residual variance by $\max\{s^2, c/\ell_I\}$. Coverage counts are simultaneous, widths are group-pair medians and interquartile ranges, and zero width is a percentage of the listed group pairs. These normal intervals are an empirical comparator without the finite-sample guarantee of the anchored CP construction.

Corpus	c	Covered	Width	Width IQR	Zero (%)	Pairs
IMDB	0	0	0.37	[0.21, 0.77]	12.1	1033
IMDB	0.25	21	0.42	[0.25, 0.77]	0.0	1033
IMDB	1	33	0.45	[0.26, 0.96]	0.0	1033
SMS	0	2	0.00	[0.00, 0.19]	52.2	728
SMS	0.25	33	0.19	[0.09, 0.69]	0.0	728
SMS	1	39	0.23	[0.12, 0.93]	0.0	728
Yelp	0	0	0.36	[0.00, 1.00]	27.0	1095
Yelp	0.25	27	0.49	[0.29, 1.00]	0.0	1095
Yelp	1	39	0.73	[0.36, 1.00]	0.0	1095
TREC	0	0	0.15	[0.00, 0.35]	42.3	742
TREC	0.25	28	0.24	[0.07, 0.54]	0.0	742
TREC	1	37	0.26	[0.12, 0.70]	0.0	742
AGNews	0	0	0.17	[0.00, 0.52]	33.8	942
AGNews	0.25	31	0.27	[0.16, 0.72]	0.0	942
AGNews	1	35	0.36	[0.19, 1.00]	0.0	942

H.3 Probability interpretation and normalization

We study label proportions within subgroups of a fixed unlabeled pool. We define this target and the existing confidence-constraint construction before explaining why group size affects interval width.

A proportion $p_I^\star(y) = 0$ means no group member has that label, whereas $p_I^\star(y) = 1$ means every member does. This target describes the observed pool, with uncertainty arising from its unknown labels. For example, a coordinate can indicate whether a labeling function voted for the candidate label y. A zero tolerance fixes the corresponding moment exactly and a larger tolerance permits more label assignments, potentially producing wider intervals.

> **Proposition 1 (restated).** If $\mathcal{U}$ is non-empty and $\mathbb{P}\{|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\} \geq 1 - \delta$, then with probability at least $1 - \delta$ the intervals contain $p_I^\star(y)$ simultaneously for all non-empty $I \subseteq \{1, \ldots, n\}$ and all $y \in \mathcal{Y}$.

The proof inserts the true point-mass assignment $p_i(y) = \mathbb{1}\{y_i = y\}$ into (1). On the confidence event this assignment is feasible, so every group proportion lies between its optimized endpoints.

The probability concerns the labeled draws and requires no independence between labeling functions. For a fixed set of constraints, the universal statement also covers groups chosen after inspecting the intervals. Section 5 addresses the different task of using labels to learn additional anchor constraints.

The pool-to-group ratio $\rho_I = n/|I|$ equals one for the full pool and increases as the group becomes smaller. Changing a label probability inside I changes its group proportion with weight $1/|I|$. Each global feature average weights that instance by $1/n$. This difference in normalization allows small-group proportions to vary substantially within the global tolerances. This ratio also appears in the endpoint dual (Álvarez et al., 2025, Eqs. (5)–(6)). The global-width lower bound quantifies the interval width imposed by this difference in scale.

I Local-constraint construction and estimation

This appendix expands the construction and local-label estimation used in Section 4. It explains the local row, confidence allocation and the two width regimes.

I.1 Local-row normalization and confidence allocation

Write the primal anchor as $|a_I^\top p - \hat{\tau}_I| \le \lambda_I$, with $(a_I)_i = \mathbb{1}\{i \in I\}/|I|$ and $p_i = p_i(1)$. Its upper and lower inequalities have nonnegative dual multipliers whose signed difference is ν_I. Eliminating their redundant common part gives the tolerance penalty $\lambda_I|\nu_I|$ in the dual. There is no ρ_I prefactor because the row already uses the group normalization $1/|I|$.

When comparing methods at the same total budget δ, we allocate part of that budget to the anchors. The global event then receives $\delta_0 < \delta$, which increases its tolerance. For example, splitting the budget evenly gives $\delta_0 = \delta/2$. For normal-quantile tolerances with $d = 12$ and $\delta = 0.1$, the global tolerance increases by 8.6%. The reference tolerance assigns the full budget to the global coordinates. This analytic normal-quantile example illustrates budget allocation; the experiments use CP calibration. Section 6 evaluates interval widths with both methods calibrated to their joint confidence events.

I.2 Binomial estimation and local sample size

Proposition 7 bounds the augmented width through the local tolerance λ_I. We next construct that tolerance from the labeled draws and determine its dependence on group size. The key quantity is the number of draws that fall inside the group.

For the uniform sampling sequence J used to estimate global moments, let $\ell_I := \sum_{r=1}^{\ell} \mathbb{1}\{J_r \in I\}$. This counts repeated positions with their multiplicities. When $\ell_I \ge 1$, let $\hat{\tau}_I$ be the fraction of these draws with label 1. Conditionally on ℓ_I, the count $\ell_I\hat{\tau}_I$ has distribution $\mathrm{Bin}(\ell_I, p_I^\star(1))$. Here $\mathrm{Bin}(a, b)$ denotes the number of successes in a independent trials with success probability b. The relation holds because an independent uniform draw conditioned to fall in I remains uniform within I.

To convert this count into a valid local constraint, we use the Clopper–Pearson interval (Clopper & Pearson, 1934). It inverts binomial tail probabilities to cover the unknown group proportion at the allocated level $1 - \delta_I$. We choose λ_I as the larger distance from $\hat{\tau}_I$ to its two endpoints. The resulting symmetric constraint contains that interval and fits (3). When $\ell_I = 0$, $(\hat{\tau}_I, \lambda_I) = (1/2, 1/2)$ gives the full-range local row. These formulas describe the symmetric text anchor and the radius used in Theorem 8. The image implementation retains the raw CP endpoints and uses their midpoint and half-width for the LP representation.

The number of group hits satisfies $\ell_I \sim \mathrm{Bin}(\ell, \rho_I^{-1})$, with mean ℓ/ρ_I. A multiplicative Chernoff bound (Hoeffding, 1963) controls the probability of receiving fewer than half this mean:

$$\mathbb{P}\{\ell_I < \ell/(2\rho_I)\} \le e^{-\ell/(8\rho_I)},$$

where $\ell/(2\rho_I)$ is the lower sample-size threshold used in the width calculation. Above this threshold we can bound the local estimation error using only ℓ and ρ_I.

The Clopper–Pearson radius is bounded by the variance-dependent expression in Theorem 8. Its square-root term reflects variation in the observed labels and its inverse-sample-size term accounts

for finite samples. The endpoint argument below establishes this connection for the radius used to construct the anchors.

I.3 Clopper–Pearson endpoint bound

We show that the symmetric radius containing the Clopper–Pearson interval satisfies the assumption of Theorem 8. Write $m = \ell_I \geq 1$ and $q = \hat{\tau}_I$. Let $t = \log(2/\delta_I)$ and $v = q(1-q)$. For either nontrivial endpoint p of the two-sided interval, the corresponding binomial tail at p equals $\delta_I/2$. The exponential bound for that tail is $\exp\{-md(q,p)\}$, where

$$d(q,p) = q\log(q/p) + (1-q)\log\{(1-q)/(1-p)\}.$$

Thus $d(q,p) \leq t/m$. The endpoint equal to q when all labels agree has distance zero and satisfies the same bound by continuity.

Put $x = |p-q|$. The identity $\partial_p d(q,p) = (p-q)/\{p(1-p)\}$ and the bound $u(1-u) \leq v + x$ between q and p imply

$$d(q,p) \geq \frac{x^2}{2(v+x)}.$$

For the upper endpoint this follows by integrating from q to p. For the lower endpoint we reverse the limits and integrate the nonnegative numerator $q-u$. Combining the inequalities yields $x^2 \leq 2(t/m)(v+x)$. Solving this quadratic and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives

$$x \leq t/m + \sqrt{(t/m)^2 + 2vt/m} \leq \sqrt{2vt/m} + 2t/m \leq \sqrt{2vt/m} + 3t/m.$$

The larger endpoint distance therefore obeys the stated bound with $A_{\delta_I} = \sqrt{2t}$ and $B_{\delta_I} = 3t$. These are the constants used in Theorem 8. The construction of the interval and its allocated confidence level are unchanged. □

I.4 Interior and boundary proportions

Remark 11 relates the two terms to the observed square-root dependence.

Remark 11 (Two regimes)**.** The transition between the terms in (4) occurs at the scale $\ell_I \asymp \log(2/\delta_I)/(\hat{p}(1-\hat{p}))$. In the *interior* regime, the group has enough labeled draws and $\hat{p}$ stays away from 0 and 1. The square-root term then dominates and the group-size exponent is $1/2$. In the *boundary* regime, few draws or an almost pure group make the B_{δ_I}/ℓ_I term dominant. This term scales linearly with ρ_I and decreases as ℓ^{-1}. Relative to the unanchored scale $\rho_I/\sqrt{\ell}$, its ratio therefore decreases as $\ell^{-1/2}$. At $n = 1600$ and $\ell = 400$, the sufficient-count probability lower bound is about 0.15 for a group of 5 instances and 0.96 for a group of 100. This count event quantifies precision and is separate from the confidence event controlling coverage. Section 6 examines both regimes.

I.5 Group-size scaling interpretation

The global construction amplifies pool estimation error by ρ_I. In contrast, anchoring estimates the group proportion from an average of ℓ/ρ_I labeled draws. The resulting leading error scales as $\sqrt{\rho_I/\ell}$. Thus more estimation labels reduce both terms in (4), while the group size determines how many of those labels reach its anchor.

I.6 Equivalent representations of the constraints

For binary feature differences $\Delta_i = \mathbf{\Phi}(\Lambda_i, 1) - \mathbf{\Phi}(\Lambda_i, 0)$, $\mathbb{E}_{p_i}\mathbf{\Phi}(\Lambda_i, y) = \mathbf{\Phi}(\Lambda_i, 0) + p_i(1)\Delta_i$ is affine in $p_i(1)$.

The intersection retains assignments compatible with both global and local information. An empty anchor family recovers exactly the construction in (1).

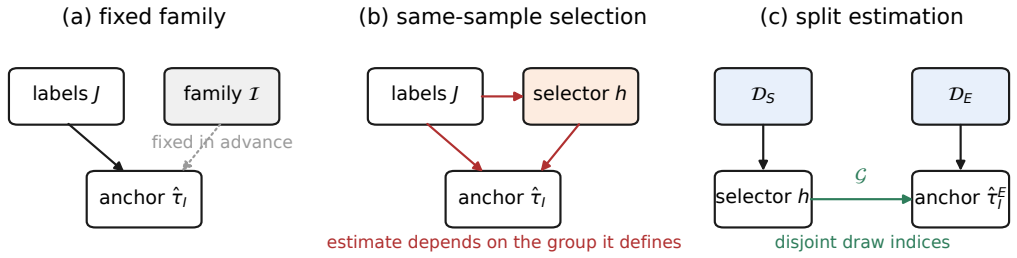


Figure A1: Where the labels enter. (a) Conditional on the fixed pool and labeling-function outputs, a pre-specified family gives a binomial anchor count conditional on its hit count. (b) Same-sample selection makes the group depend on the labels used to estimate its anchor. (c) Splitting assigns selection and estimation to disjoint draw indices, restoring the binomial relation conditionally on $\mathcal{D}_S$; a pool instance may occur in both samples.

J Learned-group construction

This appendix gives the selector construction and sample-use details for Section 5. These choices instantiate the independent-estimation requirement of Theorem 9.

J.1 Sample dependence

Figure A1 summarizes why this independence matters. Sampling with replacement still permits a pool instance to occur in both samples.

The three cases distinguish choosing a query for an existing valid set from learning new constraints using the estimation labels. Independent estimation provides independent information for the latter task.

J.2 Constructing and evaluating learned groups

Theorem 9 establishes coverage after separating selection from estimation. We now specify how to construct the learned groups and quantify the accompanying use of the label budget. The procedure consists of selection followed by confidence-constraint estimation and endpoint optimization.

Selection and group construction. To examine the effect of fitting labels at different levels of detail, we use the same labeling-function inputs for each selector. The pattern selector h_1 returns the Laplace-smoothed positive rate for each complete output pattern. It can therefore retain label information specific to individual patterns. The pooled naive Bayes selector h_2 shares class-conditional parameters across patterns. The logistic selector h_3 fits an L_2-regularized model to one-hot encoded labeling-function outputs. We set its inverse regularization strength to $C = 1$. Section 6 compares same-sample and split estimates on matched groups for each selector.

We also vary how far the scores move from the labeled prevalence. Writing $\hat{\pi}_S$ for selection-sample prevalence, the mixture is $h_w = wh + (1 - w)\hat{\pi}_S$. For a pattern u observed N_u times with K_u positive labels, the text pattern score is $(K_u + 1)/(N_u + 2)$, including $1/2$ at $N_u = 0$. The image pattern selector instead uses $\hat{\pi}_S$ for unseen patterns. For naive Bayes, $P(\Lambda_j = v \mid y = c)$ is estimated as $(N_{jvc} + 1)/(N_c + 3)$ for $v \in \{-1, 0, 1\}$. The text prior is $\hat{\pi}_S$ clipped to $[10^{-6}, 1 - 10^{-6}]$; the image prior is $(K_S + 1)/(|S| + 2)$. Posterior log-odds are clipped at magnitude 60 for text and 40 for images before the logistic transform. The logistic selector uses three-level one-hot encoding, $C = 1$, an intercept and the L-BFGS solver with at most 1000 iterations and tolerance 10^{-10}. For a single observed class it returns the selection prevalence. The endpoints $w = 1$ and $w = 0$ use the fitted score and the prevalence alone, respectively. The choice $w = 0.5$ lies halfway between them and is evaluated in Section 6. All selector settings are determined without estimation labels.

We threshold each score at nine levels from 0.90 to 0.50 in steps of 0.05. This grid controls the resolution of the positive and negative groups. We retain groups with at least 5 pool instances and remove duplicate memberships. Appendix G records these thresholds and the minimum group size. The retained groups therefore satisfy the same size and resolution criteria.

Table A5: Joint LP and local-only intervals on matched learned groups, pooling text and image settings. Failures count units with any uncovered group after the full-range solver fallback. Solver failures are listed separately. Strict narrowing counts feasible group–repetition pairs. The last column is the median joint/local width ratio over feasible pairs with positive local width; its denominator differs from the split/full ratio.

Anchor sample	Units	Local fail	Joint fail	Solver fail	Strict pairs	Joint/local
Same	3640	332	335	1	14431	1.00
Split	3640	21	25	0	15100	1.00
Full reference	3640	143	147	0	11026	1.00

Constraint estimation and optimization. After selection, let $\mathcal{I}_{\mathrm{cal}}$ be the deduplicated calibration family. In text it contains the pre-specified base families and the learned groups; in images it contains only the learned groups. Set $\delta_0 = \delta/2$ for the global event and $\delta_I = (\delta/2)/|\mathcal{I}_{\mathrm{cal}}|$ for each anchor. With no anchors, the global budget remains $\delta/2$. This equal allocation applies a union bound, with no step-down testing. Fixed global features use all acquired draws J, whereas split anchors use E only. The comparison using selection labels replaces E by S and the full reference uses J, keeping the selected family fixed.

For binary labels, the direct formulation has n probability variables and $2d + 2|\mathcal{I}_{\mathrm{cal}}|$ moment inequalities. The implementation instead aggregates the U unique labeling-function patterns. For each query it uses $2U$ positive-label mass variables, splitting each pattern into its portion inside and outside the query, bounded by those portions' counts. Anchors are unions of patterns, making this representation exact. Two endpoint solves are required per query; constraint counts alone do not determine runtime. If either endpoint solver fails to return a valid solution, the recorded interval is $[0, 1]$ with a failure flag. The fixed-group comparison also retains its existing global-box recovery: an infeasible global box is enlarged to include the labeling-function-only reference moment. This enlargement preserves the original truth-inclusion event.

Endpoints from the joint program. Table A5 compares local-only and joint LP intervals on the same selected groups, with global moments estimated from all draws. The split arm is feasible throughout and strictly narrows many group intervals. Its measured failure count is slightly higher for the joint set because intersecting valid confidence constraints controls coverage through their joint event, not through width inclusion alone. In the same-sample arm, one unit has a failed endpoint solve and receives the full range. We separate this flag from the failures of the returned intervals rather than counting it as an uncovered unit.

Precision cost. An even split leaves $\ell/2$ draws for estimation. In the square-root-dominated regime, halving this sample size gives an idealized half-width ratio of $\sqrt{2} \approx 1.414$. The realized ratio also depends on which groups are selected and how many share the local confidence budget. In addition, discrete group-hit counts and the $1/\ell_I$ correction affect small groups. These quantities explain why the idealized ratio is a reference scale.

We compare the split width with the width using all ℓ labeled draws on the same group. This full-draw reference reuses selection labels and therefore serves as a precision comparison without inheriting the independent-estimation guarantee. Section 6 also compares coverage when anchors reuse the selection labels. We summarize simultaneous coverage for each domain and repetition under each selector setting, which defines the experimental unit. A unit is covered when every group in that unit is covered. For width comparisons we divide the median split width in a unit by its median full width. The complete procedure yields intervals for learned subgroups, and the experiments in Section 6 test their coverage and width against the theoretical predictions.

K Experimental details and selector comparisons

This appendix supports the comparisons in Section 6. It gives data construction details and the selector and score-shrinkage comparisons underlying the reported coverage effects.

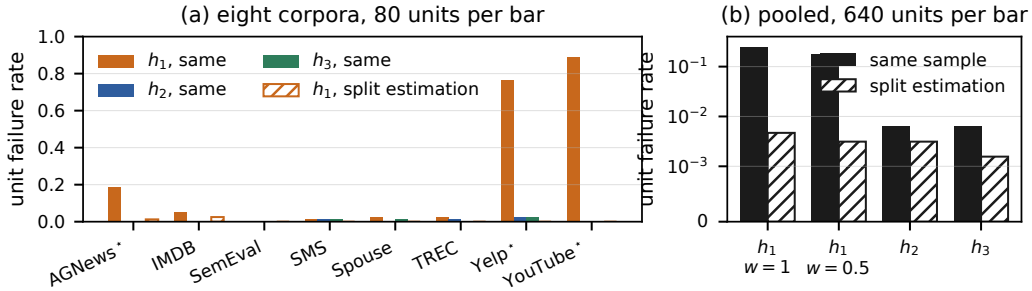


Figure A2: **Independent estimation reduces the coverage gap between selectors.** Bars show unit failure rates (↓) for local-only symmetric CP intervals on the text tasks. Panel (a) uses 80 repetitions per corpus, with solid bars for same-sample estimation and hatched bars for the split pattern selector. Stars identify the previously selected domains with concentrated same-sample failures (AGNews, Yelp and YouTube). Panel (b) pools 640 units per setting on a symmetric-log axis and compares same-sample with split estimation.

K.1 Dataset references and source construction

Image experiments use CIFAR-10N and CIFAR-100N human annotations (Wei et al., 2022; Krizhevsky, 2009) to test the same mechanism beyond text rules. The binary targets are animals against vehicles and fine-label parity. We combine human annotation streams and their agreement indicators with ridge predictors (Hoerl & Kennard, 1970) fitted to ResNet-18 features (He et al., 2016), giving weak votes that draw on both annotator responses and image features. The ridge predictors use weak labels alone. We fit them on a disjoint subset of 30,000 images and sample evaluation pools from the remaining 20,000 images, with $n = 4000$ and 60 repetitions per setting. For CIFAR-10N the labeling functions include the aggregate and worst human annotation streams alongside independent annotation rounds, while CIFAR-100N supplies fine and coarse annotation streams whose agreement contributes an additional weak signal. Clean labels define the true proportions for evaluation. The source splits and labeling-function composition are specified in Appendix G. These tasks extend the comparison to a different source of weak supervision.

The text tasks in Section 6.1 use IMDB (Maas et al., 2011) and SMS (Zhang et al., 2021) alongside Yelp (Zhang et al., 2015). TREC (Li & Roth, 2002) and AGNews (Zhang et al., 2015) supply the multiclass reductions. The learned-group comparison additionally uses SemEval (Hendrickx et al., 2010) and Spouse and YouTube from WRENCH (Zhang et al., 2021). Labeling-function selection uses unlabeled coverage (Ratner et al., 2016). The image labeling functions combine human annotations from CIFAR-10N and CIFAR-100N (Wei et al., 2022; Krizhevsky, 2009) with ridge predictors (Hoerl & Kennard, 1970) on ResNet-18 features (He et al., 2016). The complete source splits and feature composition are given in Appendix G.

K.2 Text selectors and score shrinkage

We also test whether score shrinkage removes this dependence. Mixing the pattern scores with the selection-sample prevalence at $w = 0.5$ reduces the same-sample undercoverage count to 113. The effect varies across corpora. The same-sample failures disappear on AGNews while Yelp and YouTube retain most of their failures, as shown by the paired score settings in Table A6. Shrinkage changes which groups pass a threshold without supplying independent information to estimate their proportions. Its improvement therefore leaves a different sampling problem from the one addressed by splitting. Both the fitted scores and their shrinkage target depend on selection labels, so changing the mixture alters group membership while the same-sample anchor continues to estimate prevalence from observations already used to determine that membership.

Figure A2 resolves the coverage counts in Section 6.4 into domain-level rates and compares the selector settings after pooling units across corpora. Panel (a) separates the domains where pattern-specific selection produces the largest coverage differences. Panel (b) uses a symmetric-log axis to display the same-sample and split rates for all settings.

Table A6: Local-only text intervals by corpus, 80 units per row. Counts are simultaneous-coverage failures, and inflation is split/full median-width inflation as defined in Table 3.

Corpus	Selector	Same failures	Split failures	Inflation
AGNews	Pattern, $w = 1$	15	1	1.70
AGNews	Pattern, $w = 0.5$	0	1	1.53
IMDB	Pattern, $w = 1$	4	2	1.49
IMDB	Pattern, $w = 0.5$	4	1	1.49
SemEval	Pattern, $w = 1$	0	0	1.60
SemEval	Pattern, $w = 0.5$	0	0	1.55
SMS	Pattern, $w = 1$	1	0	1.54
SMS	Pattern, $w = 0.5$	1	0	1.53
Spouse	Pattern, $w = 1$	2	0	1.63
Spouse	Pattern, $w = 0.5$	0	0	1.62
TREC	Pattern, $w = 1$	2	0	1.58
TREC	Pattern, $w = 0.5$	0	0	1.50
Yelp	Pattern, $w = 1$	61	0	1.74
Yelp	Pattern, $w = 0.5$	55	0	1.70
YouTube	Pattern, $w = 1$	71	0	1.70
YouTube	Pattern, $w = 0.5$	53	0	1.72

Table A7: Local-only text intervals by corpus, 80 units per row. Counts are simultaneous-coverage failures, and inflation is split/full median-width inflation as defined in Table 3.

Corpus	Selector	Same failures	Split failures	Inflation
AGNews	Naive Bayes	0	0	1.56
AGNews	Logistic	0	0	1.61
IMDB	Naive Bayes	0	0	1.48
IMDB	Logistic	0	0	1.49
SemEval	Naive Bayes	0	0	1.52
SemEval	Logistic	0	0	1.60
SMS	Naive Bayes	1	0	1.53
SMS	Logistic	1	0	1.58
Spouse	Naive Bayes	0	0	1.52
Spouse	Logistic	1	0	1.57
TREC	Naive Bayes	1	0	1.62
TREC	Logistic	0	0	1.55
Yelp	Naive Bayes	2	1	1.55
Yelp	Logistic	2	1	1.56
YouTube	Naive Bayes	0	1	1.62
YouTube	Logistic	0	0	1.71

K.3 Complete fixed-family and learned-group comparisons

Table A9 gives the complete fixed-family comparison supporting Section 6. Figure 2 compares image-group coverage and the width cost of splitting.

K.4 Budget and local-resolution diagnostics

Figure A3 and Table A11 trace median tightening across four label budgets using saved matched draw streams. The global reference G_1 contains the same features as the anchored construction, including pool prevalence. The prevalence-only median factor compares G_0 with G_1 on the same pairs; a median near one allows individual groups to improve.

Table A8: Uncertainty and dispersion for Table 3. The first two columns are descriptive 95% Wilson intervals for unit failure rates (percent), pooling the eight corpora. The last column is the empirical interquartile range of within-unit split/full width inflation, not a confidence interval.

Selector	Same rate CI (%)	Split rate CI (%)	Inflation IQR
Pattern, $w = 1$	[21.2, 27.8]	[0.2, 1.4]	[1.53, 1.73]
Pattern, $w = 0.5$	[14.9, 20.8]	[0.1, 1.1]	[1.49, 1.65]
Naive Bayes	[0.2, 1.6]	[0.1, 1.1]	[1.48, 1.60]
Logistic	[0.2, 1.6]	[0.0, 0.9]	[1.51, 1.66]

Table A9: Calibration ablation at $\ell = 400$, 40 repetitions per corpus. Coverage uses all queried groups; median widths use the anchored group–repetition pairs. G_0 denotes the original global feature set. Only joint CP allocates confidence across global coordinates. The coordinatewise alternatives test calibration and do not have the same simultaneous guarantee.

Corpus	Calibration	G_0 covered	A covered	G_0 width	A width
IMDB	Joint CP	40	39	1.00	0.53
IMDB	Coordinate Wilson	39	39	1.00	0.53
IMDB	Coordinate SE	32	20	1.00	0.51
SMS	Joint CP	39	40	1.00	0.75
SMS	Coordinate Wilson	27	38	1.00	0.74
SMS	Coordinate SE	12	21	0.44	0.59
Yelp	Joint CP	40	40	1.00	0.79
Yelp	Coordinate Wilson	37	38	1.00	0.79
Yelp	Coordinate SE	31	28	1.00	0.79
TREC	Joint CP	39	39	0.88	0.43
TREC	Coordinate Wilson	31	32	0.58	0.41
TREC	Coordinate SE	14	22	0.47	0.35
AGNews	Joint CP	40	40	1.00	0.43
AGNews	Coordinate Wilson	36	35	0.86	0.43
AGNews	Coordinate SE	20	14	0.57	0.41

Table A10: Coverage extension at $\ell = 400$: 100 repetitions per corpus, combining the original 40 with 60 additional saved repetitions. G_0 is the original global construction. The 95% Wilson interval describes Monte Carlo uncertainty in the anchored coverage rate (percent); it does not construct the subgroup intervals. Pair counts describe width observations, not independent coverage repetitions.

Corpus	G_0 covered	A covered	Coverage CI (%)	Group pairs
AGNews	100	100	[96.3, 100.0]	2354
IMDB	100	99	[94.6, 99.8]	2576
SMS	99	100	[96.3, 100.0]	1810
TREC	99	99	[94.6, 99.8]	1844
Yelp	100	99	[94.6, 99.8]	2736

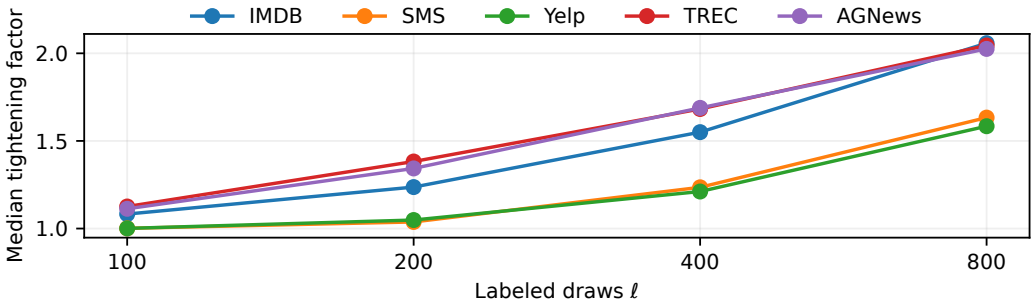


Figure A3: Median paired tightening W_{G_1}/W_A across the four tested labeled-draw budgets. Each corpus and budget uses 40 repetitions on matched pool and draw streams. These are descriptive medians over group–repetition pairs.

Table A11: Paired budget scan, 40 repetitions per corpus and budget. Tightening is median W_{G_1}/W_A, and prevalence-only is median W_{G_0}/W_{G_1} on the same anchored pairs. Coverage is the number of simultaneously covered repetitions for A. Smaller budgets are prefixes of the same draw streams; no groups are retuned for the scan.

Corpus	ℓ	Repetitions	Tightening	Prevalence-only	Covered
IMDB	100	40	1.08	1.00	40
IMDB	200	40	1.24	1.00	40
IMDB	400	40	1.55	1.00	39
IMDB	800	40	2.06	1.00	39
SMS	100	40	1.00	1.00	39
SMS	200	40	1.04	1.00	40
SMS	400	40	1.24	1.00	40
SMS	800	40	1.63	1.00	40
Yelp	100	40	1.00	1.00	40
Yelp	200	40	1.05	1.00	40
Yelp	400	40	1.21	1.00	40
Yelp	800	40	1.58	1.00	38
TREC	100	40	1.13	1.00	40
TREC	200	40	1.38	1.00	40
TREC	400	40	1.68	1.00	39
TREC	800	40	2.04	1.00	39
AGNews	100	40	1.11	1.00	40
AGNews	200	40	1.34	1.00	40
AGNews	400	40	1.69	1.00	40
AGNews	800	40	2.03	1.00	40

Table A12: Size strata at $\ell = 400$, 40 repetitions. Each factor is median W_{G_1}/W_A on anchored group–repetition pairs. The cut $|I| = 32$ is descriptive. The last two columns give median sufficient saturation sizes $ns^\star$ for original global features G_0 and prevalence-augmented features G_1; they do not define an exact onset of anchoring gains.

Corpus	Pairs < 32	Factor	Pairs ≥ 32	Factor	G_0 threshold	G_1 threshold
IMDB	334	1.07	699	2.00	19.04	19.64
SMS	459	1.12	269	1.77	25.91	26.29
Yelp	633	1.03	462	2.41	50.00	50.81
TREC	143	1.08	599	1.86	47.80	48.41
AGNews	349	1.03	593	2.13	33.65	34.30

The size-stratified comparison in Table A12 uses a descriptive cut at 32 instances. The saturation sizes are sufficient conditions for the global uncertainty sets and are reported separately for G_0 and G_1. Across the saved 30 corpus–labeling-function configurations, the share of queried groups below the size cut correlates negatively with median tightening (Spearman $r = -0.89$). This is an association across configurations sharing five corpora, with the share counting groups rather than pool instances.

Table A13 separates fits to the local tolerance from fits to final LP width and records how often the clipped local constraint determines that width. Additional reductions are conditional summaries over strictly narrowed pairs, not average improvements over all pairs. Table A14 separately verifies fixed-tolerance monotonicity and sufficient saturation on saved endpoints.

Table A13: Local-precision diagnostics. Slopes fit log median tolerance λ_I or joint width W_A against log group size within the listed observed size range (20 repetitions, equal weight per distinct size), excluding medians at or within 10^{-9} of zero or one. Binding means equality with the clipped local-only width to tolerance 10^{-7}, over all feasible pairs in the 40-repetition batch. Extra reduction is a median percentage only among the listed strictly narrowed pairs.

Corpus	Size range	λ slope	R^2	W_A slope	Binding (%)	Strict pairs	Extra (%)
IMDB	25–1600	-0.50	0.99	-0.41	90.1	102	3.1
SMS	25–1600	-0.61	0.99	-0.44	87.8	89	4.5
Yelp	25–1600	-0.48	0.98	-0.36	91.8	90	4.8
TREC	71–1600	-0.55	0.99	-0.53	81.0	141	4.3
AGNews	61–1600	-0.48	0.97	-0.34	90.1	93	7.6

Table A14: Constraint checks at $\ell = 400$, 40 repetitions. Widening compares A with G_1 on all listed anchored pairs, either sharing total δ or keeping global tolerances fixed. The last columns count queried groups below their own repetition-specific G_0 saturation threshold and how many lack a full-range G_0 interval. All comparisons use tolerance 10^{-7}.

Corpus	Anchor pairs	Wider: split δ	Wider: fixed λ	Below threshold	Non-full
IMDB	1033	40	0	220	0
SMS	728	46	0	418	0
Yelp	1095	40	0	663	0
TREC	742	49	0	160	0
AGNews	942	39	0	353	0

L SCOPE OF THE INTERVAL GUARANTEES

This appendix specifies the scope of the interval guarantees in Sections 4 and 5, including selector dependence and the finite-population range of the local estimation bound.

The theory covers groups that are fixed or independent of their estimation labels. Theorem 9 covers adaptive selection by using independent estimation labels. Splitting imposes no restriction on selector complexity. The empirical selector comparison concerns three constructions at one fixed threshold resolution. A selector that memorizes labels for individual labeling-function patterns has higher undercoverage than the two pooled selectors tested. This comparison does not establish coverage for pooled selectors in general.

Splitting supplies independent estimation labels while halving the sample available to each anchor. Small groups incur both the $\sqrt{2}$ scale and the discreteness correction. The median image width cost is 1.42 to 1.98. Local hit counts determine how this cost varies across selected groups. Theorem 8 relates this cost to the resolution of the queried groups. The finite-population lower bound supports the asymptotic $1/2$ scale and is vacuous at the evaluated sample-to-group ratios. The upper and lower constants remain unmatched. The image extension uses a fixed binary task on each corpus. It does not evaluate full ten-class or hundred-class interval estimation.