Partial Identification from LLM Prompts

Xiaohong Chen
Yale University

Ashesh Rambachan
MIT

Elie Tamer
Harvard University

June 24, 2026

Abstract

Large language models are increasingly used as binary classifiers when the true label is latent. We study partial identification of the prevalence ($\theta = P(X^* = 1)$) from panels of LLM reports whose errors may be arbitrarily dependent given the truth. The design of replication determines the observable, and hence the identifying content: repeated prompts to one model yield a count, several named models a response vector, and both a response matrix. Cast as a two-component finite mixture, the problem makes the identification failure transparent—absent restrictions that separate the latent components, the prevalence θ is completely unidentified, and weak stochastic-ordering restrictions (first-order dominance, monotone likelihood ratio, mean ordering) leave the identified set at $[0, 1]$. Identifying power comes instead from externally calibrated scores and events, which discipline the mixture in the spirit of the misclassification and corrupted-data literature. We characterize the resulting bounds, establishing validity and sharpness, and give an exact account of the identifying information in the full score distribution beyond its mean. When named models are asked repeated versions of the same question, what identifies θ is not the number of positive answers but which models agree across prompts—a feature a vote count discards. An extension derives implied bounds on regression coefficients when X^* is a regressor of interest that is not directly observed.

1 Introduction

Large language models (LLMs) are now routinely used as binary classifiers. They label text as toxic or non-toxic, factual or non-factual, policy-violating or safe, relevant or irrelevant, and so on. In many applications the target label is not observed in the main sample. The econometrician observes only LLM reports and wants to learn the latent prevalence

$$\theta = \mathbb{P}(X^* = 1),$$

where $X^* \in \{0, 1\}$ is the true label.

The central problem is not only that LLMs make mistakes. It is that there is more than one way to replicate an LLM measurement, and different replications produce different observable objects. This paper distinguishes three designs given in Table 1 below.

Table 1: Three LLM measurement-panel designs

Design	Data for one item	Recommended observable	Reason
One LLM, repeated questions or prompt variants	$R_1, \ldots, R_M$	Count $S = \sum_{m=1}^{M} R_m$	No model identity to preserve; if prompt variants are exchangeable, prompt labels carry no structural content.
Many named LLMs, one question each	$Y = (Y_1, \ldots, Y_J)$	Full named vector $Y \in \{0, 1\}^J$	Model identities matter: LLMs differ in sensitivity, specificity, refusal behavior, and bias.
Many named LLMs, repeated questions each	$R = (R_{jm})_{j \le J, m \le M}$	Full matrix R; under prompt exchangeability, the column-pattern histogram N (lossless); otherwise the model-count vector $T_j = \sum_m R_{jm}$ as a practical coarsening	The matrix is sharpest. N preserves named-model agreement within exchangeable prompts; T preserves model identity while aggregating prompt variation.

Notation reserves J for named models and M for repeated prompts. The first row covers the common "same LLM asked J repeated questions" case after relabeling that count as M.

If the same LLM is asked repeated exchangeable versions of the same binary question, a count is natural. By contrast, if GPT-4, GPT-4 Turbo, Claude, and open-weight models are each asked once, replacing the named vector by a vote count throws away information: the response patterns $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$ have the same count but different implications if the first two models are more reliable. If each named model is asked

repeated prompt variants, neither the simple count nor the single-prompt named vector is adequate; the analyst observes a two-way measurement panel and must decide whether model identity, prompt identity, or both should be preserved.

Paper Contributions The baseline nonidentification result and the use of sensitivity or specificity-style restrictions to bound a prevalence are applications of partial identification techniques (see Manski (1999), Tamer (2010)). The paper's contribution is to adapt, organize, and extend that literature to handle LLM measurement panels, where errors are plausibly dependent across models and prompts, and to add results that are, to our knowledge, new in this setting:

1. A design taxonomy (Table 1) that maps the source of replication—repeated prompts, named models, or both—into the correct observable object, with an exact characterization of when a coarsening of the response matrix is *truth-sufficient* (Proposition 5).
2. A unified calibration theory: all of our identifying restrictions are calibrated scores or calibrated events. We prove validity of the resulting bounds, give an exact *sharpness characterization* of the score bounds via a rearrangement function of the observed score distribution (Theorem 3), show the simple linear bounds are sharp when only the score mean is recorded and exactly sharp for binary scores (Corollary 1), and prove sharpness of the event bounds.
3. A symmetry result for the two-way design: under prompt exchangeability and an explicit restriction-compatibility condition (Assumption 1), the full matrix has a lossless reduction to the column-pattern histogram N (Theorem 4). No independence is assumed anywhere.
4. Diagnostics—coarsening loss, row/column influence, dependence audits, and a minimum-tolerance specification test—that add transparency without adding identifying assumptions, together with a multiple-testing-aware treatment of optimized calibrated events.

An augmented empirical illustration quantifies the practical payoff: in a toxicity-labeling application with three named LLMs, reporter-specific calibration on the named vector roughly halves the width of the identified set relative to the count coarsening used in earlier drafts.

2 Relation to the literature

We connect our results to important literatures.

Latent-class and multiple-rater models. Estimating rater error rates without a gold standard goes back at least to Dawid and Skene (1979). That tradition typically obtains point identification through conditional independence of raters given the truth (or low-order dependence corrections). Our setting deliberately drops conditional independence: LLMs share training corpora, benchmarks, synthetic data, distillation pipelines, and alignment procedures, so their errors can be arbitrarily dependent given X^*. Without independence, the Dawid–Skene identification route is unavailable, and the model becomes a two-component mixture with unrestricted components—hence partial identification.

Partial Identification with Misclassification, corrupted data and Mixtures. Bounds on parameters under misclassified or contaminated outcomes are classical (Horowitz and Manski, 1995; Molinari, 2008; Hu, 2008). Our calibrated-score bounds are recognizably of this family: bounds on sensitivity and specificity translate linearly into bounds on prevalence. What we add is (i) the score/event organization that nests reporter-specific accuracy, thresholds, unanimity, weighted ensembles, and matrix events as one assumption type rather than many; and (ii) the sharpness analysis of Section 4, which distinguishes what is sharp given the score mean from what is sharp given the score law. Finally Henry et al. (2014) study partial identification of finite mixtures using observable variation in mixture weights. Our degeneracy result (Proposition 1) is the mixture-identification observation specialized to LLM panels; its value is the discipline it imposes on applied work, not mathematical novelty. Our group-level extension (Appendix A) connects directly to the exclusion-restriction logic of that literature.

Classifier calibration and LLM-based labeling. A growing literature calibrates classifier and LLM confidence and uses calibrated predictions for prevalence estimation (Silva Filho et al., 2023; Hovsepian et al., 2024; Linder et al., 2026). We treat external calibration as the *source of identification*: validated lower confidence bounds on score sensitivity/specificity or event predictive values are exactly the inputs our bounds require. The division of labor is deliberate: that literature supplies calibrated constants; this paper says what those constants identify under arbitrary dependence.

3 Framework and nonidentification

3.1 Latent truth, response matrices, and summaries

For each item, let $X^* \in \{0,1\}$ be the latent truth. Let $j = 1, \ldots, J$ index named LLMs and $m = 1, \ldots, M$ index repeated questions, prompt variants, stochastic completions, or elicitation templates. The binary response from model j under prompt m is $R_{jm} \in \{0,1\}$, and the full response matrix is

$$R = (R_{jm})_{j \le J,\, m \le M} \in \{0,1\}^{J \times M}.$$

The parameter of interest is $\theta = \mathbb{P}(X^* = 1)$. In the spirit of the partial identification literature, we develop bounds on θ using minimal plausible assumptions.

Write $\pi_R(r) = \mathbb{P}(R = r)$ and, conditional on the latent state, $f_z(r) = \mathbb{P}(R = r \mid X^* = z)$ for $z \in \{0, 1\}$. Then

$$\pi_R(r) = (1 - \theta) f_0(r) + \theta f_1(r), \qquad r \in \{0, 1\}^{J \times M}. \tag{1}$$

No factorization of f_z is assumed: entries of R may be arbitrarily dependent within each latent state.

The analyst may work with a finite summary $U = g(R) \in \mathcal{U}$: a count, a named response vector, a model-count vector, a prompt-count vector, a column-pattern histogram, or the matrix itself. Let $p_U(u) = \mathbb{P}(U = u)$ and $q_{z,U}(u) = \mathbb{P}(U = u \mid X^* = z)$. Every summary satisfies

$$p_U(u) = (1 - \theta) q_{0,U}(u) + \theta q_{1,U}(u), \qquad u \in \mathcal{U}. \tag{2}$$

For restrictions $\mathcal{A}_U$ on $(q_{0,U}, q_{1,U})$, define the identified set

$$\Theta_U(p_U; \mathcal{A}_U) = \big\{\theta \in [0, 1] : \exists\, (q_{0,U}, q_{1,U}) \text{ satisfying (2) and } \mathcal{A}_U\big\}. \tag{3}$$

3.2 Degeneracy and weak ordering

Proposition 1 (Nonidentification of θ)**.** *Fix any finite summary $U = g(R)$. Without restrictions that rule out equality of the latent component distributions, $\Theta_U(p_U) = [0, 1]$: for every $\theta \in [0, 1]$, the choice $q_{0,U} = q_{1,U} = p_U$ satisfies (2) and the data contain no information about θ.*

Proof. For any u, $(1 - \theta) p_U(u) + \theta p_U(u) = p_U(u)$, for every $\theta \in [0, 1]$. □

Weak shape restrictions do not alter this conclusion. For count summaries, first-order stochastic dominance (FOSD), monotone likelihood ratio (MLR) ordering, and weak mean ordering formalize the idea that positive items should produce more positive reports; for vector or matrix summaries, coordinatewise stochastic orders play the same role. These restrictions are often plausible, but they permit $q_0 = q_1$, so the degenerate decomposition remains feasible.

For clarity, in a count experiment $S \in \{0, \ldots, M\}$, FOSD means

$$\sum_{t=s}^{M} q_1(t) \geq \sum_{t=s}^{M} q_0(t), \qquad s = 1, \ldots, M.$$

MLR means that q_1/q_0 is increasing in the usual cross-product sense: for $s > s'$, $q_1(s)q_0(s') \geq q_1(s')q_0(s)$, with the standard conventions at zeros. Weak mean ordering means $\mathbb{E}[S \mid X^* = 1] \geq \mathbb{E}[S \mid X^* = 0]$. For vector or matrix summaries, coordinatewise FOSD means $\mathbb{E}[\varphi(U) \mid X^* = 1] \geq \mathbb{E}[\varphi(U) \mid X^* = 0]$ for every bounded coordinatewise increasing function φ; equivalently, the inequality holds for all increasing upper sets. All of these are weak orders. Hence $q_0 = q_1 = p_U$ satisfies them with equality.

3.3 A confidence-set implication

The nonidentification result has an inferential counterpart. Let $\mathcal{M}_U$ be a class of joint laws for (X^*, U). Suppose that for every observable law $p \in \Delta(\mathcal{U})$ and every t in a set $\Theta_0 \subseteq [0,1]$, the uninformative experiment $U \sim p$, $X^* \sim \text{Bernoulli}(t)$, $X^* \perp U$ belongs to $\mathcal{M}_U$.

Theorem 1 (Distribution-free impossibility)**.** *Let $\widehat{\Theta}_{n,\alpha} \subseteq [0,1]$ be a possibly randomized confidence set for θ, constructed from an i.i.d. sample $U_1, \ldots, U_n$. If $\sup_{P \in \mathcal{M}_U} \mathbb{P}_P\{\theta(P) \notin \widehat{\Theta}_{n,\alpha}\} \leq \alpha$, then for every observable law p and every $t \in \Theta_0$, $\mathbb{P}_p\{t \notin \widehat{\Theta}_{n,\alpha}\} \leq \alpha$. Consequently*

$$\mathbb{E}_p\big[\lambda(\widehat{\Theta}_{n,\alpha} \cap \Theta_0)\big] \geq (1-\alpha)\lambda(\Theta_0),$$

where λ is Lebesgue measure. If $\Theta_0 = [0,1]$ then $\mathbb{E}_p[\lambda(\widehat{\Theta}_{n,\alpha})] \geq 1-\alpha$, and if the confidence set is always an interval contained in $[0,1]$ then $\mathbb{P}_p\{\widehat{\Theta}_{n,\alpha} = [0,1]\} \geq 1-2\alpha$.

Proof. Fix p and $t \in \Theta_0$. The law with $U \sim p$, $X^* \sim \text{Bernoulli}(t)$, $X^* \perp U$ belongs to $\mathcal{M}_U$; under it the sample has distribution p^n and the prevalence is t, so coverage implies $\mathbb{P}_p\{t \notin \widehat{\Theta}_{n,\alpha}\} \leq \alpha$. Integrating over $t \in \Theta_0$ (Tonelli) gives the expected-length bound. If $\Theta_0 = [0,1]$ and the set is always an interval, coverage of both endpoints implies the interval is $[0,1]$; the final claim follows by the union bound. □

3.4 Coarsening and information loss

Theorem 2 (Coarsening weakens identification under compatible restrictions)**.** *Let $U_2 = h(U_1)$. Suppose every feasible $(\theta, q_{0,U_1}, q_{1,U_1})$ under restrictions $\mathcal{A}_{U_1}$ projects to a feasible $(\theta, q_{0,U_2}, q_{1,U_2})$ under restrictions $\mathcal{A}_{U_2}$. Then $\Theta_{U_1}(p_{U_1}; \mathcal{A}_{U_1}) \subseteq \Theta_{U_2}(p_{U_2}; \mathcal{A}_{U_2})$.*

Proof. Given a feasible point under U_1, define $q_{z,U_2}(u_2) = \sum_{u_1 : h(u_1) = u_2} q_{z,U_1}(u_1)$. Aggregating (2) over fibers of h gives the mixture equation for U_2; compatibility gives the projected restrictions. □

Section 7 sharpens this in the matrix design: Proposition 5 characterizes exactly when a coarsening is lossless for the truth, and Theorem 4 gives a design symmetry under which a specific coarsening is lossless for the identified set.

4 Identification by calibration

This section contains the paper's maintained identifying content. Everything else—reporter-specific accuracy, threshold classifiers, relaxed support, unanimity, weighted votes, matrix-agreement rules—is a special case obtained by choosing the summary U, the score w, or the events A, B.

4.1 Calibrated score bounds: validity

Let $w : \mathcal{U} \to [0,1]$ be a pre-specified score and write

$$\bar{w} = \mathbb{E}[w(U)] = \sum_{u \in \mathcal{U}} w(u)\, p_U(u).$$

The score may be a count fraction, a threshold rule, a named-model report, a weighted ensemble, or a matrix-agreement statistic.

Proposition 2 (Calibrated score bounds)**.** *Suppose*

$$\mathbb{E}[w(U) \mid X^* = 1] \geq a, \qquad \mathbb{E}[1 - w(U) \mid X^* = 0] \geq b, \qquad a, b \in (0,1]. \tag{4}$$

Then every admissible prevalence satisfies

$$\max\Big\{0,\ \frac{\bar{w} + b - 1}{b}\Big\} \ \leq\ \theta \ \leq\ \min\Big\{1,\ \frac{\bar{w}}{a}\Big\}. \tag{5}$$

Proof. Let $\mu_z = \mathbb{E}[w(U) \mid X^* = z]$, so $\bar{w} = (1-\theta)\mu_0 + \theta\mu_1$. Since $\mu_1 \geq a$ and $\mu_0 \geq 0$, $\bar{w} \geq \theta a$, giving the upper bound. Since $\mu_0 \leq 1 - b$ and $\mu_1 \leq 1$, $\bar{w} \leq (1-\theta)(1-b) + \theta = 1 - b + b\theta$, giving the lower bound. Intersect with $[0,1]$. □

4.2 Sharpness: score mean vs score law

The interval (5) uses only the *mean* $\bar{w}$ of the score. When the analyst observes the full law of $w(U)$, the sharp identified set can be strictly smaller, and it admits an exact characterization through a rearrangement (concentration) function.

For $t \in [0,1]$ define

$$W^+(t) \ =\ \max\Big\{\textstyle\sum_u w(u)h(u)\ :\ 0 \leq h(u) \leq p_U(u)\ \forall u,\ \ \sum_u h(u) = t\Big\}. \tag{6}$$

$W^+(t)$ is the largest possible contribution to $\mathbb{E}[w(U)]$ from a sub-population of mass t: it is computed by a greedy fill, allocating mass to the values of u with the largest $w(u)$ first (a Hardy–Littlewood rearrangement bound). W^+ is concave and nondecreasing, with $W^+(0) = 0$ and $W^+(1) = \bar{w}$, and satisfies $W^+(t) \le \min\{t, \bar{w}\}$.

Theorem 3 (Sharp identified set under score calibration)**.** *Maintain* (4) *and suppose the analyst observes* p_U *(hence the law of* $w(U)$*) but imposes no other restriction. The identified set for* θ *is*

$$\Theta_w \;=\; \Big\{\theta \in [0,1] \;:\; W^+(\theta) \;\ge\; a\theta \;\; \textit{and} \;\; W^+(\theta) \;\ge\; \bar{w} - (1-b)(1-\theta)\Big\}, \tag{7}$$

and Θ_w *is a (possibly empty) closed interval.*

Proof. Work with the joint mass $h_1(u) = \mathbb{P}(X^* = 1, U = u)$ and $h_0 = p_U - h_1$. Feasibility of a prevalence θ is equivalent to the existence of h_1 with

$$0 \le h_1 \le p_U, \qquad \sum_u h_1(u) = \theta, \qquad \sum_u w(u)h_1(u) \ge a\theta, \qquad \sum_u w(u)h_0(u) \le (1-b)(1-\theta),$$

where the third inequality is $\mathbb{E}[w\,\mathbf{1}\{X^* = 1\}] \ge a\,\mathbb{P}(X^* = 1)$, i.e. (4) for μ_1, and the fourth is (4) for μ_0. The fourth inequality rewrites as $\sum_u w(u)h_1(u) \ge \bar{w} - (1-b)(1-\theta)$. The feasible set for h_1 given the first two constraints is a nonempty polytope (for $\theta \in [0,1]$), and the achievable values of the linear functional $s = \sum_u w(u)h_1(u)$ over that polytope form a closed interval $[W^-(\theta), W^+(\theta)]$. Hence a feasible h_1 exists if and only if $W^+(\theta) \ge \max\{a\theta,\ \bar{w} - (1-b)(1-\theta)\}$. Both $\theta \mapsto a\theta$ and $\theta \mapsto \bar{w} - (1-b)(1-\theta)$ are affine and W^+ is concave, so each constraint defines an interval and Θ_w is their intersection. □

Corollary 1 (Relation to the linear bounds; exact sharpness for binary scores)**.** (i) Θ_w *is contained in the interval* (5). (ii) *If* $w(U) \in \{0,1\}$ *almost surely, then* Θ_w *equals* (5)*: the linear bounds are exactly sharp for binary scores (in particular, for all event indicators and threshold classifiers).* (iii) *The interval* (5) *is sharp in the class of observed laws with score mean* $\bar{w}$*: for every* θ *in* (5) *there exists a law of* $w(U)$ *with mean* $\bar{w}$ *and a feasible decomposition supporting* θ*. Hence* (5) *cannot be improved using* $\bar{w}$ *alone.*

Proof. (i) From $W^+(\theta) \le \bar{w}$ and $W^+(\theta) \ge a\theta$ we get $\theta \le \bar{w}/a$; from $W^+(\theta) \le \theta$ and $W^+(\theta) \ge \bar{w} - (1-b)(1-\theta)$ we get $b\theta \ge \bar{w} + b - 1$. (ii) For binary w, $W^+(\theta) = \min\{\theta, \bar{w}\}$. If $\theta \le \bar{w}$ the first constraint in (7) holds since $a \le 1$, and the second reduces to $\theta \ge (\bar{w} + b - 1)/b$; if $\theta > \bar{w}$ the second holds automatically and the first reduces to $\theta \le \bar{w}/a$. Since $(\bar{w} + b - 1)/b \le \bar{w} \le \bar{w}/a$, the union of the two regimes is exactly (5). (iii) Given $\bar{w}$, take $w(U)$ binary with $\mathbb{P}(w(U) = 1) = \bar{w}$ and apply (ii). □

Remark 1 (Interpretation). The gap between Θ_w and (5) is an exact measure of the information in the *shape* of the score distribution beyond its mean. For binary scores

the shape carries nothing extra; for graded scores (e.g. $w = S/M$ with M large) the rearrangement constraint $W^+(\theta) \geq a\theta$ can bind strictly earlier than $\theta = \bar{w}/a$, tightening the upper bound. W^+ is computed by sorting, so the sharp set costs no more than the linear bounds in practice (Section 8). Emptiness of Θ_w is a specification rejection of the calibration constants (a, b) at the observed law.

4.3 Calibrated events: validity and sharpness

Let $A \subseteq \mathcal{U}$ be a high-confidence positive event and $B \subseteq \mathcal{U}$ a high-confidence negative event.

Proposition 3 (Posterior event calibration; sharp)**.** *If $\mathbb{P}(X^* = 1 \mid U \in A) \geq \rho$ then $\theta \geq \rho\,\mathbb{P}(U \in A)$. If $\mathbb{P}(X^* = 0 \mid U \in B) \geq \lambda$ then $\theta \leq 1 - \lambda\,\mathbb{P}(U \in B)$. Each one-sided bound is sharp under the corresponding single event-calibration restriction.*

Proof. The lower bound follows from

$$\theta \geq \mathbb{P}(X^* = 1, U \in A) = \mathbb{P}(X^* = 1 \mid U \in A)\mathbb{P}(U \in A) \geq \rho\mathbb{P}(U \in A),$$

and the upper bound is analogous. For sharpness of the lower bound, under the single restriction involving A, set $h_1(u) = \rho p_U(u)$ for $u \in A$ and $h_1(u) = 0$ otherwise, with $h_0 = p_U - h_1$. Then $0 \leq h_1 \leq p_U$, the calibration constraint holds with equality whenever $\mathbb{P}(U \in A) > 0$, and $\theta = \sum_u h_1(u) = \rho\mathbb{P}(U \in A)$. The upper bound is symmetric, assigning mass $h_0(u) = \lambda p_U(u)$ on B and $h_0(u) = 0$ off B. □

Proposition 4 (Wrong-state event errors; sharp)**.** *If $\mathbb{P}(U \in A \mid X^* = 0) \leq \alpha_A$ with $\alpha_A \in [0, 1)$, then*

$$\theta \;\geq\; \max\Big\{0,\; \frac{\mathbb{P}(U \in A) - \alpha_A}{1 - \alpha_A}\Big\}.$$

If $\mathbb{P}(U \in B \mid X^* = 1) \leq \alpha_B$ with $\alpha_B \in [0, 1)$, then

$$\theta \;\leq\; \min\Big\{1,\; \frac{1 - \mathbb{P}(U \in B)}{1 - \alpha_B}\Big\}.$$

Each one-sided bound is sharp under the corresponding single wrong-state error restriction.

Proof. Let $p_A = \mathbb{P}(U \in A)$. Since $p_A \leq (1 - \theta)\alpha_A + \theta$, rearrangement gives the lower bound. To see sharpness, if $p_A \leq \alpha_A$, the bound is zero and is attained by $\theta = 0$ with $q_0 = p_U$. If $p_A > \alpha_A$, set

$$\theta_A = \frac{p_A - \alpha_A}{1 - \alpha_A}.$$

Allocate positive-state joint mass only inside A, proportionally to p_U on A, with total mass θ_A, and set $h_0 = p_U - h_1$. Then $h_0(A) = p_A - \theta_A = \alpha_A(1 - \theta_A)$, so $\mathbb{P}(U \in A \mid$

$X^* = 0) = \alpha_A$ and the lower bound is attained. The upper bound is symmetric: if $\mathbb{P}(U \in B) \leq \alpha_B$, take $\theta = 1$; otherwise set $\theta_B = (1 - \mathbb{P}(U \in B))/(1 - \alpha_B)$, allocate all mass outside B to the positive state and allocate additional positive-state mass inside B so that $\mathbb{P}(U \in B \mid X^* = 1) = \alpha_B$. □

Remark 2 (One source of identification). Propositions 2–4 are the maintained identifying content of the paper. Reporter-specific accuracy, threshold classifiers, relaxed support, unanimity, weighted votes, and matrix-agreement rules are not separate assumptions; they are obtained by choosing different U, w, A, and B. Proposition 4 is the one-sided binary-score analogue of Proposition 2: the lower bound uses $w = \mathbf{1}\{U \in A\}$ and the specificity-type constraint $\mathbb{E}[1 - w(U) \mid X^* = 0] \geq 1 - \alpha_A$, while the upper bound uses $w = \mathbf{1}\{U \in B\}$ and the false-negative constraint $\mathbb{E}[w(U) \mid X^* = 1] \leq \alpha_B$. The direct construction above gives sharpness.

4.4 Optimized calibrated events and multiplicity

Rich summaries admit many candidate high-agreement events. Rather than choosing one arbitrarily, the analyst can pre-specify a finite class and calibrate the whole class on a validation sample, treating event selection explicitly as a multiple-testing problem.

Let $\mathcal{A}^+$ be a finite class of positive events $A \subseteq \mathcal{U}$ (row-threshold events, column-pattern events, trusted-model events, weighted-score threshold events). Suppose validation data deliver simultaneous lower confidence bounds $\widehat{\rho}_A$ such that, with probability at least $1-\alpha$,

$$\mathbb{P}(X^* = 1 \mid U \in A) \;\geq\; \widehat{\rho}_A \quad \text{for every } A \in \mathcal{A}^+. \tag{8}$$

Then with the same probability all lower bounds hold simultaneously, so

$$\theta \;\geq\; \max_{A \in \mathcal{A}^+} \; \widehat{\rho}_A \, \mathbb{P}(U \in A), \tag{9}$$

and symmetrically $\theta \leq \min_{B \in \mathcal{A}^-} \{1 - \widehat{\lambda}_B \mathbb{P}(U \in B)\}$ for a simultaneously calibrated negative class $\mathcal{A}^-$.

Remark 3 (Pre-specification and honest calibration). The event class must be fixed before examining the main unlabeled sample, or the calibration step must account for selection. Simultaneity in (8) can be obtained by Bonferroni corrections, split-sample validation, or conformal-style calibration. The same discipline applies to the calibration constants (a_j, b_j) used anywhere in the paper: for honest inference they should be *lower confidence bounds* estimated on a sample (or sample split) disjoint from the one used to compute observed rates. This adds no structural assumption; it only ensures the validity statements survive the search over events.

4.5 Optional sensitivity restrictions

The paper's baseline identification comes from calibrated scores and calibrated events. Two additional restrictions are useful as sensitivity analyses, but they should not be presented as maintained assumptions unless independently justified. For a count or vote score $C \in \{0, \ldots, J\}$, directional asymmetry with parameter $\gamma > 0$ imposes

$$\mathbb{E}[C \mid X^* = 0] \leq \gamma \, \mathbb{E}[J - C \mid X^* = 1],$$

so false-positive votes are bounded relative to false-negative votes. For a one-LLM count $S \in \{0, \ldots, M\}$ with observed mean $\bar{s} = \mathbb{E}[S]$, anchored separation with tolerance $\varepsilon > 0$ imposes

$$\mathbb{E}[S \mid X^* = 1] \geq \bar{s} + \varepsilon, \qquad \mathbb{E}[S \mid X^* = 0] \leq \bar{s} - \varepsilon,$$

which implies $\theta \geq \varepsilon/(\varepsilon + M - \bar{s})$ and $\theta \leq \bar{s}/(\bar{s} + \varepsilon)$. These restrictions are informative because they impose cross-state separation, not because they use replication by itself.

Table 2: Restrictions and their role in the paper

Role	Restriction or object	Status
Regularity	First-order stochastic dominance (FOSD), monotone likelihood ratio (MLR), weak mean ordering	Plausible but non-identifying; optional background restrictions; definitions appear in Section 3.
Main identifying content	Calibrated score $w(U)$	Core result (Proposition 2, Theorem 3); calibrated from validation data or external evidence.
Main identifying content	Calibrated event $A \subseteq \mathcal{U}$ or $B \subseteq \mathcal{U}$	Core result (Propositions 3–4); includes high- and low-agreement regions; sharp.
Design structure	Exchangeability of prompt labels, when built into the design	Not an accuracy assumption; justifies lossless reduction from R to the orbit statistic N (Theorem 4).
Special cases	Reporter-specific accuracy, threshold rules, relaxed support, unanimity, weighted votes	Instances of score or event calibration, not separate assumptions.
Optional sensitivity	Directional asymmetry, anchored separation, conditional independence	Not baseline; define explicitly, use only when substantively justified, and report separately.
Diagnostics	Coarsening loss, row/column influence, dependence audits, minimum tolerance	Empirical transparency tools; add no identifying power.

5 Design I: one LLM, repeated binary questions

A single LLM is used repeatedly to measure the same latent binary truth. For one item, write the repeated reports as $R_1, \ldots, R_M \in \{0, 1\}$. If the repeated prompts are exchangeable probes of the same truth, their labels carry no structural content and the natural observable is the count

$$S = \sum_{m=1}^{M} R_m \in \{0, \ldots, M\}, \qquad p(s) = (1-\theta)q_0(s) + \theta q_1(s).$$

No conditional independence is imposed across prompts: repeated prompts from one model may share the same systematic errors.

Remark 4 (When the count is appropriate). The count is appropriate only if the repetitions are exchangeable measurements of the same latent truth. If prompts ask substantively different questions, there is no single scalar X^* behind all responses. If prompt identities have known reliability differences, keep the prompt-response vector rather than counting.

The count design has two natural calibrated scores: $w_1(S) = S/M$ and $w_k(S) = \mathbf{1}\{S \geq k\}$.

Average repeated-prompt accuracy. With $w = S/M$ in Proposition 2 and $\bar{r} = \mathbb{E}[S/M]$,

$$\max\left\{0, \frac{\bar{r} + b - 1}{b}\right\} \leq \theta \leq \min\left\{1, \frac{\bar{r}}{a}\right\}. \tag{10}$$

Because S/M is a graded score, Theorem 3 applies with content: the sharp set computed from W^+ can be strictly inside (10), and is obtained by sorting the support of S.

Threshold accuracy. For a threshold k, let $D_k = \mathbf{1}\{S \geq k\}$ and $p_k^+ = \mathbb{P}(S \geq k)$. If validation data support $\mathbb{P}(D_k = 1 \mid X^* = 1) \geq a_k$ and $\mathbb{P}(D_k = 0 \mid X^* = 0) \geq b_k$, then

$$\max\left\{0, \frac{p_k^+ + b_k - 1}{b_k}\right\} \leq \theta \leq \min\left\{1, \frac{p_k^+}{a_k}\right\},$$

and by Corollary 1(ii) these bounds are exactly sharp.

High- and low-count events. With $A = \{S \geq k\}$ and $B = \{S \leq \ell\}$, Proposition 3 gives $\theta \geq \rho_k \mathbb{P}(S \geq k)$ and $\theta \leq 1 - \lambda_\ell \mathbb{P}(S \leq \ell)$; Proposition 4 gives the tail-error bounds

$$\theta \;\geq\; \max\left\{0, \frac{\mathbb{P}(S \geq k) - \alpha_{0k}}{1 - \alpha_{0k}}\right\}, \qquad \theta \;\leq\; \min\left\{1, \frac{1 - \mathbb{P}(S \leq \ell)}{1 - \alpha_{1\ell}}\right\}.$$

Relaxed support is the special case $k = M$, $\ell = 0$: $q_0(M) \leq \alpha_0$, $q_1(0) \leq \alpha_1$. Exact support ($\alpha_0 = \alpha_1 = 0$) gives $p(M) \leq \theta \leq 1 - p(0)$ but is rarely credible for LLMs; relaxed, validation-calibrated tail bounds are usually preferable.

Implication. Design I is useful only to the extent that the count distribution can be calibrated. Report a small number of validation-calibrated score or event bounds—ideally the sharp set of Theorem 3 for the graded score—rather than many sensitivity restrictions.

6 Design II: many named LLMs, one binary question each

J named LLMs each answer the same binary question once. The response vector is $Y = (Y_1, \ldots, Y_J) \in \{0,1\}^J$ with law $\pi(y)$ and class-conditionals $f_z(y)$ satisfying $\pi(y) = (1-\theta)f_0(y) + \theta f_1(y)$. No conditional independence is assumed across named LLMs. The vote count $\sum_j Y_j$ is a coarsening, not the primitive object.

The named-vector design matters because it permits calibrated scores and events that use model identity: the single-reporter score $w_j(Y) = Y_j$, the weighted score $w(Y) = \sum_j c_j Y_j$ with $c_j \geq 0$, $\sum_j c_j = 1$, and events $A \subseteq \{0,1\}^J$ encoding agreement by a trusted subset rather than simple majority.

Reporter-specific accuracy. Suppose validation data provide

$$\mathbb{P}(Y_j = 1 \mid X^* = 1) \geq a_j, \qquad \mathbb{P}(Y_j = 0 \mid X^* = 0) \geq b_j, \qquad j = 1, \ldots, J, \tag{11}$$

and let $m_j = \mathbb{P}(Y_j = 1)$. Applying Proposition 2 to each $w_j(Y) = Y_j$ and intersecting,

$$\max_{1 \leq j \leq J} \max\Big\{0, \frac{m_j + b_j - 1}{b_j}\Big\} \;\leq\; \theta \;\leq\; \min_{1 \leq j \leq J} \min\Big\{1, \frac{m_j}{a_j}\Big\}. \tag{12}$$

Each one-reporter bound is sharp (binary score); the intersection is valid and is the sharp set based on the marginals alone. Using the joint law π with the restrictions (11) simultaneously can tighten further; this is the LP of Section 8.

Named events. If $A \subseteq \{0,1\}^J$ satisfies $\mathbb{P}(X^* = 1 \mid Y \in A) \geq \rho_A$ then $\theta \geq \rho_A \pi(A)$; if B satisfies $\mathbb{P}(X^* = 0 \mid Y \in B) \geq \lambda_B$ then $\theta \leq 1 - \lambda_B \pi(B)$. Relaxed unanimity is only the special case $A = \{\mathbf{1}_J\}$, $B = \{\mathbf{0}_J\}$ and should not be the default event unless validation evidence supports it.

The exact cost of counting votes. If only the vote count $S = \sum_j Y_j$ is stored, the reporter-specific marginals and named events are not recoverable; the count identifies only $\bar{m} = \frac{1}{J}\mathbb{E}[S] = \frac{1}{J}\sum_j m_j$. With averaged calibration constants $\beta_1 = \frac{1}{J}\sum_j a_j$, $\beta_0 = \frac{1}{J}\sum_j b_j$, the count-only analogue is

$$\max\Big\{0, \frac{\bar{m} + \beta_0 - 1}{\beta_0}\Big\} \;\leq\; \theta \;\leq\; \min\Big\{1, \frac{\bar{m}}{\beta_1}\Big\}, \tag{13}$$

which is generally strictly weaker than (12); Section 10 quantifies the gap in the toxicity application. The same loss applies to weighted scores and trusted-subset events.

Implication. Analyze Design II with the named vector. Count-based analysis is a robustness check or a fallback when only counts were stored. The identifying content should come from reporter-specific or named-event calibration, not from generic independence assumptions.

7 Design III: many named LLMs, repeated questions each

The third design is a two-way measurement panel: for one item, observe $R = (R_{jm}) \in \{0,1\}^{J\times M}$ with mixture law (1). No conditional independence is assumed across rows, columns, or cells. Design III is valuable not because it makes independence credible, but because it records *where* agreement occurs.

7.1 Summaries and coarsenings

Useful lower-dimensional summaries: the model-count vector $T = (T_1, \ldots, T_J)$, $T_j = \sum_m R_{jm}$; the prompt-count vector $S = (S_1, \ldots, S_M)$, $S_m = \sum_j R_{jm}$; the total count $C = \sum_{j,m} R_{jm}$; and, when prompt labels are exchangeable but model labels are not, the *column-pattern histogram*

$$N_y(R) = \sum_{m=1}^{M} \mathbf{1}\{R_{\cdot m} = y\}, \qquad y \in \{0,1\}^J, \qquad N(R) = (N_y(R) : y \in \{0,1\}^J), \tag{14}$$

which counts the prompts on which the named response pattern equals y (with $J = 3$, N_{110} counts prompts where models 1 and 2 say one and model 3 says zero). The histogram preserves cross-model agreement within prompts while discarding prompt labels, and refines the model-count vector via $T_j = \sum_y y_j N_y$. The coarsening hierarchy is

$$R \longrightarrow N \longrightarrow T \longrightarrow C, \qquad R \longrightarrow S \longrightarrow C,$$

with support sizes 2^{JM}, $\binom{M+2^J-1}{2^J-1}$, $(M+1)^J$, $(J+1)^M$, and $JM+1$ respectively. The full matrix distinguishes patterns with the same total count: one weak model saying yes on every prompt is not equivalent to several trusted models each saying yes repeatedly.

7.2 Truth-sufficient reductions of the response matrix

Which coarsenings lose information about X^*? The answer is a likelihood-ratio invariance condition.

Let $\Omega = \{0,1\}^{J\times M}$, let $U = h(R)$ be any finite summary with fibers $F_u = \{r : h(r) = u\}$, and let $q_z(u) = \sum_{r\in F_u} f_z(r)$.

Proposition 5 (Truth-sufficient coarsenings)**.** *Assume* $0 < \mathbb{P}(X^* = 1) < 1$. *The following are equivalent, up to null sets.*

1. *U is sufficient for the latent truth:* $\mathbb{P}(X^* = 1 \mid R = r) = \mathbb{P}(X^* = 1 \mid U = h(r))$ *for all r with positive probability.*
2. *The residual matrix information given the summary is independent of the truth:* $\mathbb{P}(R = r \mid U = u, X^* = 1) = \mathbb{P}(R = r \mid U = u, X^* = 0)$ *for $r \in F_u$.*
3. *The likelihood ratio $L(r) = f_1(r)/f_0(r)$ is constant on each fiber F_u (usual conventions for zeros).*
4. *There is a Markov kernel $K(r \mid u)$, independent of z, with $f_z(r) = q_z(h(r))\, K(r \mid h(r))$ for $z = 0, 1$.*

When these conditions hold, R and U contain the same information about X^*; when they fail, the coarsening discards information about the latent truth.

Proof. (2)⇔(4): take $K(r \mid u) = \mathbb{P}(R = r \mid U = u, X^* = z)$, which is independent of z exactly when (2) holds. (2) is equivalent to $f_1(r)/q_1(u) = f_0(r)/q_0(u)$ on F_u, i.e. to $f_1(r)/f_0(r) = q_1(u)/q_0(u)$ constant on F_u, which is (3). By Bayes' rule, $\mathbb{P}(X^* = 1 \mid R = r) = \theta f_1(r)/\{(1-\theta)f_0(r) + \theta f_1(r)\}$ depends on r only through $h(r)$ iff f_1/f_0 does, giving (1)⇔(3). □

Remark 5 (Common matrix summaries). C is truth-sufficient only if f_1/f_0 is constant across matrices with the same total count; T only if constant across matrices with the same row-count vector; N only if constant across matrices with the same pattern histogram. Prompt exchangeability of both f_0 and f_1 is sufficient for the last property but not necessary: likelihood-ratio invariance within prompt-permutation orbits is enough.

7.3 Prompt exchangeability and a lossless matrix reduction

Some reductions are not losses: they remove labels that have no design meaning. In Design III, prompt labels may be exchangeable even when model labels are not.

Let S_M be the group of permutations of prompt labels; for $\sigma \in S_M$, $(\sigma r)_{jm} = r_{j,\sigma(m)}$. The histogram $N(R)$ indexes the orbits of this action: two matrices have the same N iff one is obtained from the other by permuting prompt columns. The action extends to distributions by $(\sigma q)(r) = q(\sigma^{-1}r)$.

The reduction from R to N is lossless only when prompt exchangeability is part of the maintained design. To avoid ambiguity, let $\Theta_R^{\mathrm{ex}}(\pi_R; \mathcal{A}_R)$ denote the full-matrix

identified set when admissible component distributions q_0, q_1 are required to be prompt-exchangeable in addition to satisfying the matrix-level restrictions $\mathcal{A}_R$. The required correspondence between matrix-level and histogram-level restrictions is explicit below.

Assumption 1 (Restriction compatibility)**.** (i) *The matrix-level restrictions $\mathcal{A}_R$ are invariant to prompt relabeling: if $(q_0, q_1) \in \mathcal{A}_R$ and the q_z are prompt-exchangeable, then relabeling prompts does not change the truth of the restrictions.* (ii) *The histogram-level restriction set is exactly the projection of the exchangeable part of $\mathcal{A}_R$:*

$$\mathcal{A}_N = \Big\{(q_0^N, q_1^N) : \ (q_0, q_1) \in \mathcal{A}_R, \ q_z(\sigma r) = q_z(r) \ \forall \sigma \in S_M, \ q_z^N \text{ is the law of } N \text{ under } q_z\Big\}.$$

Theorem 4 (Lossless reduction under prompt exchangeability)**.** *Suppose prompt exchangeability is a maintained design restriction, the observed population law π_R is prompt-exchangeable, and Assumption 1 holds. Then the full matrix and the column-pattern histogram give the same identified set:*

$$\Theta_R^{\text{ex}}(\pi_R; \mathcal{A}_R) \ = \ \Theta_N(p_N; \mathcal{A}_N).$$

Proof. ($\subseteq$) Let (θ, q_0, q_1) be feasible for $\Theta_R^{\text{ex}}(\pi_R; \mathcal{A}_R)$. Projecting each q_z along N gives (q_0^N, q_1^N) satisfying the histogram mixture equation because projection commutes with mixing. Assumption 1(ii) gives $(q_0^N, q_1^N) \in \mathcal{A}_N$.

($\supseteq$) Let (θ, q_0^N, q_1^N) be feasible for $\Theta_N(p_N; \mathcal{A}_N)$. For each histogram value n, lift $q_z^N(n)$ uniformly over the finite orbit $\{r : N(r) = n\}$. The lifted q_z are prompt-exchangeable and project back to q_z^N. Because π_R is prompt-exchangeable, it is itself the uniform orbit lift of p_N, so the lifted distributions satisfy the matrix mixture equation for π_R. Assumption 1(ii) ensures that the lifted pair satisfies $\mathcal{A}_R$. Thus the same θ is feasible for $\Theta_R^{\text{ex}}(\pi_R; \mathcal{A}_R)$. □

Remark 6 (No independence is used). The theorem uses only a design symmetry and matched restrictions. It does not assume prompts are independent conditional on X^*, nor that cells are weakly correlated; entries of R may be arbitrarily dependent within each latent state.

Remark 7 (Why compatibility matters). Both directions of Assumption 1 have bite. If the analyst imposes *additional* restrictions after reducing to N (so $\mathcal{A}_N$ is strictly smaller than the projection), then only $\Theta_N \subseteq \Theta_R^{\text{ex}}$ is guaranteed. Conversely, a matrix-level restriction that is *not* prompt-invariant—for example, a calibrated accuracy bound for prompt $m = 1$ only—cannot be expressed through N at all; dropping it in the reduction gives only $\Theta_R^{\text{ex}} \subseteq \Theta_N$. Equality is a statement about matched restriction classes, not about the statistic alone.

Remark 8 (General orbit reduction). The argument applies to any finite group of design symmetries: if G acts on the matrix space, admissible q_z are G-invariant, and the restrictions are G-invariant and matched as in Assumption 1, the orbit statistic gives the same identified set as the full matrix. Prompt exchangeability gives N; independent prompt relabeling within each model gives T; full exchangeability of all cells gives C, but that last symmetry is rarely credible for named LLMs.

7.4 Matrix scores and matrix events

Design III should not be analyzed as a large vote count unless both dimensions are intentionally treated as exchangeable. The natural calibrated objects are matrix scores and matrix events.

A weighted matrix score has the form

$$w(R) = \frac{\sum_j \sum_m c_{jm} R_{jm}}{\sum_j \sum_m c_{jm}}, \qquad c_{jm} \geq 0, \quad \sum_{j,m} c_{jm} > 0,$$

with special cases $w = T_j/M$, $w = S_m/J$, and $w = C/(JM)$; if (4) holds for w, Proposition 2 and Theorem 3 apply. Under prompt exchangeability a score may be written as a function of N, e.g. placing weight on column patterns where trusted models agree.

Matrix events encode repeated agreement across both dimensions, e.g.

$$A^+(K,L) = \Big\{R : \sum_{j=1}^{J} \mathbf{1}\{T_j \geq L\} \ \geq K\Big\}, \qquad A^-(K,L) = \Big\{R : \sum_{j=1}^{J} \mathbf{1}\{T_j \leq L\} \ \geq K\Big\},$$

the events that at least K named models each produce at least (at most) L positive prompt responses. The histogram also suggests events invisible from T: with a trusted subset $J_0 \subseteq \{1, \ldots, J\}$,

$$\sum_{y:\, y_j = 1 \ \forall j \in J_0} N_y \ \geq \ L$$

requires the trusted models to agree positively on at least L prompts, regardless of the other models, while remaining invariant to prompt relabeling. Calibrated predictive values or wrong-state errors for any of these events feed Propositions 3–4; searches over event classes use Section 4.4.

7.5 Model-specific and prompt-specific bounds; structured calibration

The model-count vector is a tractable compromise when R or N is too large. If for each named model j

$$\mathbb{E}[T_j/M \mid X^* = 1] \geq a_j, \qquad \mathbb{E}[(M - T_j)/M \mid X^* = 0] \geq b_j,$$

then with $r_j = \mathbb{E}[T_j/M]$,

$$\max_{1\le j\le J} \max\Big\{0, \frac{r_j + b_j - 1}{b_j}\Big\} \;\le\; \theta \;\le\; \min_{1\le j\le J} \min\Big\{1, \frac{r_j}{a_j}\Big\},$$

and symmetrically for calibrated prompt-count scores S_m/J. Model-specific bounds suit stable, calibratable model error profiles; prompt-specific bounds suit stable prompt families. If prompt labels are exchangeable but within-prompt agreement matters, N is preferable to T because it retains more agreement geometry.

Full cell-level calibration of every pair (j, m) may be too parameter-rich. A practical compromise groups models and prompts: with model groups $g(j)$ and prompt families $h(m)$, calibrate $\mathbb{P}(R_{jm} = 1 \mid X^* = 1) \ge a_{g(j),h(m)}$ and $\mathbb{P}(R_{jm} = 0 \mid X^* = 0) \ge b_{g(j),h(m)}$, with the grouping fixed before analysis or justified by validation data.

7.6 Diagnostics unique to the matrix design

Design III allows diagnostics that add no identifying assumptions.

Coarsening loss. For $U \in \{R, N, T, S, C\}$ and compatible restrictions, report widths of identified sets and their differences, e.g.

$$\text{Loss}(N \to T) = \text{wid}(\Theta_T) - \text{wid}(\Theta_N), \qquad \text{Loss}(T \to C) = \text{wid}(\Theta_C) - \text{wid}(\Theta_T).$$

A large loss from N to T means within-prompt cross-model agreement matters; from T to C, that model identity matters. If prompt exchangeability is credible and restrictions are invariant and matched, Theorem 4 predicts no loss from R to N at the population level—a checkable implication of the design symmetry.

Row and column influence. Let $\Theta^{(-j)}$ and $\Theta^{(-m)}$ be the identified sets after dropping model j or prompt m. Large movements in either bound show that conclusions hinge on a particular model or prompt; this is often more informative than another structural assumption.

Dependence and redundancy. On a validation sample, estimate within-model dependence $\text{Corr}(R_{jm}, R_{jm'} \mid X^* = z)$ and across-model dependence $\text{Corr}(R_{jm}, R_{j'm} \mid X^* = z)$. High correlations indicate redundant measurements; low correlations indicate nonredundant variation. These are diagnostics, not independence assumptions.

8 Computation and inference

Linear programming. For any finite summary U, fixed-θ feasibility is a linear program. With joint masses $h_z(u) = \mathbb{P}(X^* = z, U = u)$:

$$h_0(u) + h_1(u) = p_U(u), \qquad \sum_u h_1(u) = \theta, \qquad h_z \geq 0,$$

and conditional restrictions become linear after multiplying through, e.g. $\mathbb{E}[w(U) \mid X^* = 1] \geq a$ becomes $\sum_u w(u)h_1(u) \geq a \sum_u h_1(u)$. Sharp bounds under the baseline linear score and event restrictions are two LPs: minimize and maximize $\sum_u h_1(u)$. For fixed θ, FOSD restrictions are linear in the conditional component probabilities and can be included in fixed-θ feasibility checks; exact MLR restrictions are bilinear in the components and should be handled only through explicit relaxations or nonlinear optimization. Prompt exchangeability is implemented either by using N directly or by orbit-equality constraints on matrix probabilities.

The sharp score set by sorting. $W^+(t)$ in (6) is computed greedily: order support points by descending $w(u)$ and fill mass to total t. W^+ is piecewise linear and concave, so Θ_w in (7) is found by intersecting a concave piecewise-linear function with two affine functions—no LP solver needed.

Sampling uncertainty. When p_U is estimated from N items, replace mixture equalities by bands $|(1-\theta)q_0(u) + \theta q_1(u) - \widehat{p}_U(u)| \leq \Delta_u$. A simple finite-sample choice is the Hoeffding-union bound

$$\varepsilon_N(\alpha) = \sqrt{\frac{\log(2|\mathcal{U}|/\alpha)}{2N}}, \tag{15}$$

with $\Delta_u = \varepsilon_N(\alpha)$ for every u, yielding a conservative outer confidence set; less conservative alternatives use the multinomial bootstrap, empirical Bernstein bands, or moment-inequality methods. For optimized event bounds, sampling uncertainty enters twice—through $\widehat{p}_U$ in the main sample and through validation estimates of predictive values—and validity requires the simultaneous calibration (8).

Specification testing by minimum tolerance. Let $\mathcal{Q}$ be the set of distributions over $\mathcal{U}$ implied by the maintained restrictions for some θ, and define

$$\Delta^* = \min_{q \in \mathcal{Q}} \|\widehat{p}_U - q\|_\infty,$$

the minimum ℓ_∞ tolerance restoring feasibility. If the restrictions are correct at the population law p_U^0, then $\Delta^* \leq \|\widehat{p}_U - p_U^0\|_\infty$, so a finite-sample test rejects at level α if

$\Delta^* > \varepsilon_N(\alpha)$, and

$$\Delta_0 \in \left[\max\{0, \Delta^* - \varepsilon_N(\alpha)\},\ \Delta^* + \varepsilon_N(\alpha)\right]$$

is a confidence interval for the population misspecification distance. The test detects out-of-distribution collapse: if an LLM panel becomes uninformative, restrictions calibrated in distribution may become infeasible. Emptiness of the sharp score set Θ_w is the same idea specialized to a single calibrated score.

9 Simulation: count-based Beta-Binomial experiment

Set $M = 5$ and $\theta_0 = 0.65$, with class-conditional counts $S \mid X^* = 1 \sim \text{BetaBin}(5, 4, 1.5)$ and $S \mid X^* = 0 \sim \text{BetaBin}(5, 1.5, 4)$; the observed histogram is $p = (1 - \theta_0)q_0 + \theta_0 q_1$. The design has $\mathbb{E}[S \mid X^* = 0] = 1.36$, $\mathbb{E}[S \mid X^* = 1] = 3.64$, average repeated-prompt accuracy 0.727, and observed mean $\mathbb{E}[S] = 2.84$, so $\bar{w} = \mathbb{E}[S/M] = 0.568$ for the graded score $w = S/M$. Without restrictions, or under FOSD/MLR/weak mean ordering alone, the identified set is $[0, 1]$ because the degenerate decomposition $q_0 = q_1 = p$ remains feasible (Proposition 1); Table 4 records this once as a benchmark and then isolates the paper's main computational comparison: the mean-only linear bounds (5) versus the sharp rearrangement set of Theorem 3, computed from W^+ by sorting the six support points of S.

Table 3: Simulation DGP: class-conditional and mixture PMFs

s	$q_0(s)$	$q_1(s)$	$p(s)$
0	0.310	0.015	0.118
1	0.291	0.055	0.137
2	0.208	0.121	0.152
3	0.121	0.208	0.178
4	0.055	0.291	0.208
5	0.015	0.310	0.207

The table mirrors the paper's identification message in miniature. Weak ordering restrictions leave $[0, 1]$ untouched, so they are recorded in a single benchmark row. Calibration is what moves the set, and the shape of the score distribution carries identifying content beyond its mean, exactly as Theorem 3 predicts: at $a = b = 0.60$ the rearrangement constraint $W^+(\theta) \geq \bar{w} - (1-b)(1-\theta)$ binds before the linear lower bound, raising θ_L from 0.280 to 0.317; at $a = b = 0.70$ the sharp set $[0.479, 0.784]$ removes 29% of the linear interval's width (0.429 to 0.305), tightening both endpoints. Because $w(S)$ is a graded score on six support points, this gain is obtained by a sort and costs nothing computationally. Exact support tightens only the weakly calibrated case ($a = b = 0.60$, where it

Table 4: Simulation identified sets for θ: linear (mean-only) versus sharp (full-law) score bounds

Restriction	Uses	θ_L	θ_U	Width	Covers θ_0
None, or FOSD/MLR/mean order	–	≈ 0	≈ 1	1.000	Yes
Exact support, $p(5) \le \theta \le 1 - p(0)$	tail cells	0.207	0.882	0.675	Yes
Linear score (5), $a = b = 0.60$	mean $\bar{w}$	0.280	0.947	0.667	Yes
Sharp score (Thm. 3), $a = b = 0.60$	law of $w(S)$	0.317	0.947	0.630	Yes
Linear score (5), $a = b = 0.70$	mean $\bar{w}$	0.383	0.812	0.429	Yes
Sharp score (Thm. 3), $a = b = 0.70$	law of $w(S)$	0.479	0.784	0.305	Yes
Sharp score + exact support, $a = b = 0.60$	law + support	0.317	0.882	0.565	Yes
Sharp score + exact support, $a = b = 0.70$	law + support	0.479	0.784	0.305	Yes

Notes: the score is $w = S/M$ with $\bar{w} = 0.568$. Sharp sets are computed from the rearrangement function W^+ of (6) by sorting; the combined sharp-score-plus-support rows are computed by the LP of Section 8 and agree with the closed form of Theorem 3 when support does not bind.

caps θ_U at 0.882) and is redundant once calibration is strong; this ordering—calibration first, support as a benchmark—is the recommended reporting style of Section 12.

10 Empirical illustration: toxicity classification

This section revisits the toxicity-response classification task studied by Cheng et al. (2024). The task is to bound the prevalence of toxic responses when the analyst observes binary labels from LLM annotators and, for a validation sample, expert-adjudicated truth (which gives us a benchmark to allow us to validate our bounds).

In particular, the data involve $J = 3$ named LLM annotators—GPT-4, GPT-4 Turbo, and Claude-2—and three human annotators. This is a Design II setting. Following the taxonomy of this paper, we now report the named-vector analysis in three increasingly disciplined steps—plug-in marginal bounds, the full named-vector LP on the joint law of $Y = (Y_1, Y_2, Y_3)$, and an honest version that replaces plug-in calibration constants by one-sided lower confidence bounds—and we tag every bound with the response object that must be stored to compute it.

Table 5 shows why named-vector analysis matters: GPT-4 is much more accurate than Claude-2, and all three LLMs have higher specificity than sensitivity. A count-only analysis collapses this heterogeneity into an exchangeable average. The validation class-conditionals also show that exact support is empirically violated with only three LLMs: among truly toxic items, 14.2% receive zero positive LLM votes; among truly non-toxic items, 4.3% receive unanimous positive votes. Exact support can be reported as a benchmark, but relaxed support or validation-calibrated score bounds are more credible.

Table 6 records the observable objects that the bounds below are built from. The

Table 5: Annotator characteristics on validation set, $N = 1{,}000$, $\theta_0 = 0.650$

Annotator	Positive rate	Sensitivity	Specificity	Accuracy
Human 1	0.536	0.786	0.929	0.836
Human 2	0.627	0.857	0.800	0.837
Human 3	0.601	0.852	0.866	0.857
GPT-4	0.562	0.795	0.871	0.822
GPT-4 Turbo	0.483	0.662	0.849	0.727
Claude-2	0.272	0.345	0.863	0.526
LLM exchangeable average	–	0.601	0.861	–

top panel gives the count histograms—$\widehat{p}(s)$ with the validation class-conditionals $\widehat{q}_z(s)$ for $N = 1{,}000$, and the count-only training histogram for $N = 28{,}194$—while the lower panel reports the full named-vector law $\widehat{\pi}(y)$ over $\{0,1\}^3$. The two are not interchangeable: the count is the coarsening $s = \sum_j y_j$ of $\widehat{\pi}$, so patterns that disagree on *which* model flagged the item are merged. Here $(1,1,0)$, where GPT-4 and GPT-4 Turbo flag and Claude-2 abstains, has mass 0.250, whereas the equal-count pattern $(0,1,1)$ has mass only 0.010; counting pools them into $\widehat{p}(2) = 0.295$ and discards exactly the reporter identity that the named-vector bounds will exploit.

Table 6: Observed response objects, $J = 3$ LLMs: vote-count histograms and named-vector law

Validation set, $N = 1{,}000$				Training set, $N = 28{,}194$	
s	$\widehat{p}(s)$	$\widehat{q}_0(s)$	$\widehat{q}_1(s)$	s	Train $\widehat{p}(s)$
0	0.348	0.731	0.142	0	0.218
1	0.172	0.163	0.177	1	0.162
2	0.295	0.063	0.420	2	0.177
3	0.185	0.043	0.262	3	0.444
$\bar{p}$	1.317	–	–	$\bar{p}$	1.847
$\mathbb{E}[S \mid X^*]$	–	0.417	1.802		

Named-vector law $\widehat{\pi}(y)$, $y = (y_{GPT\text{-}4}, y_{Turbo}, y_{Cl\text{-}2})$, validation:

y	000	001	010	011	100	101	110	111
$\widehat{\pi}(y)$	0.348	0.042	0.038	0.010	0.092	0.035	0.250	0.185

Notes: the count histogram $\widehat{p}(s)$ is the coarsening of $\widehat{\pi}(y)$ along $s = \sum_j y_j$; the patterns $(1,1,0)$ and, e.g., $(0,1,1)$ are merged by counting even though they involve reporters of very different accuracy. For the training set only the count histogram was stored.

10.1 Named-vector bounds: plug-in benchmark and honest calibration

Table 7 reports reporter-specific bounds (12) in two versions. Panel A is the *plug-in benchmark*: calibration constants (a_j, b_j) are the validation sensitivities and specificities of Table 5, estimated on the same sample as the observed rates m_j, so the panel illustrates the identification logic but is not honest inference. Panel B is the *honest calibration* version recommended in Section 4.4: (a_j, b_j) are replaced by one-sided Clopper–Pearson lower confidence bounds (a_j^L, b_j^L) at Bonferroni level $\alpha/6$ with $\alpha = 0.05$, so all six constraints hold simultaneously with probability at least 0.95 and the resulting bounds are valid by Proposition 2.

Table 7: Reporter-specific bounds, validation set: plug-in benchmark versus honest calibration

Score	m_j	a_j	b_j	θ_L	θ_U
Panel A: plug-in benchmark (constants and rates from the same sample)					
GPT-4 ($w = Y_1$)	0.562	0.795	0.871	0.497	0.707
GPT-4 Turbo ($w = Y_2$)	0.483	0.662	0.849	0.391	0.730
Claude-2 ($w = Y_3$)	0.272	0.345	0.863	0.156	0.789
Intersection				**0.497**	**0.707**
Panel B: honest calibration (one-sided 95% simultaneous lower confidence bounds)					
GPT-4 ($w = Y_1$)	0.562	0.755	0.823	0.468	0.744
GPT-4 Turbo ($w = Y_2$)	0.483	0.615	0.797	0.351	0.785
Claude-2 ($w = Y_3$)	0.272	0.300	0.813	0.105	0.906
Intersection				**0.468**	**0.744**

Notes: bounds use (12). In Panel B, a_j^L and b_j^L are exact one-sided Clopper–Pearson lower confidence bounds for sensitivity ($n_1 = 650$) and specificity ($n_0 = 350$) at level 0.05/6 each. Both panels cover $\theta_0 = 0.650$; honest calibration widens the intersection from $[0.497, 0.707]$ to $[0.468, 0.744]$ (width 0.209 to 0.277)—the price of validity is modest. Fully honest inference would additionally compute (a_j^L, b_j^L) on a split disjoint from the sample used for m_j and $\widehat{\pi}$, and add sampling bands (15).

10.2 The key comparison: what storing richer objects buys

Table 8 is the section's central exhibit. It reports four identified sets for the same population, ordered by the richness of the stored response object, including the full named-vector LP of Section 8, which imposes the mixture equation on the joint law $\widehat{\pi}(y)$ over $\{0,1\}^3$ together with all reporter-specific calibration restrictions simultaneously.

Three facts stand out. First, the cost of counting: moving from the named vector to the count widens the set from $[0.497, 0.707]$ to $[0.348, 0.731]$—the width nearly doubles, from 0.209 to 0.383—purely because counting discards reporter identity, quantifying Theorem 2. Second, the full LP on the joint law coincides with the marginal intersection

Table 8: Key comparison: identified sets for θ by stored response object, validation set

Bound	Object stored	θ_L	θ_U	Width
Count coarsening, eq. (13)	count S	0.348	0.731	0.383
Reporter-specific marginal intersection	named vector Y (marginals)	0.497	0.707	0.209
Full named-vector LP	joint law of Y	0.497	0.707	0.209
Honest full named-vector LP	joint law of Y	0.468	0.744	0.277

Notes: all rows cover $\theta_0 = 0.650$. Row 1 uses the exchangeable-average constants $(\beta_1, \beta_0) = (0.601, 0.861)$ with $\bar{m} = 0.439$. Rows 2–3 use the plug-in constants of Table 7, Panel A; row 4 uses the honest constants of Panel B. The honest count-coarsened analogue of row 1 (averaging the a_j^L, b_j^L) is $[0.308, 0.788]$, width 0.480: the storage gain survives honest calibration, 0.277 versus 0.480.

to four decimals. This is informative rather than disappointing: it certifies that, given reporter-specific calibration alone, the marginal intersection is already sharp at this observed law, so no further information can be extracted from these restrictions; the joint law is nevertheless the object to store, because only it makes the sharpness check possible and because pattern-level calibrated events (trusted-subset agreement, Section 4.4) require it whenever validation evidence supports them. Third, honest calibration costs about 0.07 of width relative to the plug-in benchmark but preserves the storage ranking: the honest named-vector set (width 0.277) remains far narrower than the honest count set (width 0.480).

The binding reporter on both sides is GPT-4, the most accurate model; the influence diagnostic of Section 7 applies verbatim: dropping GPT-4 widens the plug-in intersection to the Turbo bounds $[0.391, 0.730]$. For the training set (N = 28,194), only the count histogram was stored, so every named-vector row of Table 8 is unavailable by construction: with $\bar{m} = 1.847/3 = 0.616$ and the exchangeable-average constants, (13) gives [0.554, 1.000]. The storage choice has real consequences: on the training set the count bound is $[0.554, 1.000][0.554, 1.000][0.554, 1.000]$, whose upper endpoint is the trivial value 1, so counting leaves the prevalence bounded only from below. Had the named vector been stored, both endpoints would be informative.

11 Extension: regression on a latent label

So far the target has been the scalar prevalence $\theta = \mathbb{P}(X^* = 1)$. In many applications X^* is not the final object but a latent *regressor*: the analyst wants the partial association between an observed outcome and the latent label, holding covariates fixed. This section shows that the calibration bounds feed directly into such a regression, with prevalence recovered as the special case of a constant outcome.

For each item let (V, Z, U) be observed, where $V \in \mathbb{R}$ is a scalar outcome, $Z \in \mathbb{R}^d$ a

vector of observed covariates, and $U = g(R)$ the LLM summary of Sections 3–7; $X^* \in \{0,1\}$ remains latent. (We write the outcome as V to avoid collision with the named-LLM vector Y of Section 6.) Stack the regressors as $W = (1, Z', X^*)' \in \mathbb{R}^{d+2}$ and consider the best linear predictor (BLP) coefficient

$$\beta \;=\; \big(\mathbb{E}[WW']\big)^{-1}\mathbb{E}[WV],$$

assuming $\mathbb{E}[WW']$ is nonsingular. The coefficient on X^*, denoted γ, is typically the parameter of interest; if $\mathbb{E}[V \mid Z, X^*]$ is linear, β is the conditional-expectation coefficient.

Reduction to latent cross-moments. Because (V, Z) are observed and X^* is binary (so $(X^*)^2 = X^*$), every entry of $\mathbb{E}[WW']$ and $\mathbb{E}[WV]$ is identified from the data except the blocks that pair X^* with an observed variable:

$$\theta = \mathbb{E}[X^*], \qquad \mathbb{E}[X^*Z] \in \mathbb{R}^d, \qquad \mathbb{E}[X^*V] \in \mathbb{R}.$$

Collect these in $\mu = (\theta, \mathbb{E}[X^*Z]', \mathbb{E}[X^*V])'$. Then $\beta = \Phi(\mu; m_{\text{obs}})$ for a known map Φ that is rational in μ—a matrix inverse times a vector, both affine in μ—where m_{obs} collects the observed moments. Identifying β thus reduces to identifying μ.

Bounding the latent cross-moments. Since (U, Z, V) are jointly observed, write the unknown attachment of the latent label to the data as

$$\eta(u, z, v) = \mathbb{P}(X^* = 1 \mid U = u, Z = z, V = v) \in [0, 1].$$

Every latent cross-moment is a *linear* functional of η: for any observed h,

$$\mathbb{E}[X^*\, h(U, Z, V)] = \mathbb{E}[h(U, Z, V)\, \eta(U, Z, V)].$$

Lemma 1 (Calibrated bounds on latent moments)**.** *Let h be any observed, bounded function. Over all η consistent with the observed law of (U, Z, V) and with the maintained calibration restrictions of Section 4, the cross-moment $\mathbb{E}[X^*h]$ ranges over a closed interval whose endpoints solve the linear programs*

$$\min_{\eta} \;/\; \max_{\eta} \; \mathbb{E}[h(U, Z, V)\, \eta(U, Z, V)] \quad \textit{s.t.} \quad 0 \le \eta \le 1, \; \textit{marginal consistency, calibration.}$$

Taking $h \equiv 1$ returns the prevalence bounds of Section 4 exactly; taking $h \in \{Z_1, \ldots, Z_d, V\}$ bounds the remaining components of μ.

The calibration inequalities enter exactly as in Section 8: a score restriction $\mathbb{E}[w(U) \mid X^* = 1] \ge a$ becomes $\sum_u w(u)\, \mathbb{E}[\eta \mid U = u]\, p_U(u) \ge a\, \mathbb{E}[\eta]$ after writing $\mathbb{E}[X^* w(U)] = \mathbb{E}[w(U)\eta]$, and an event restriction becomes the corresponding linear inequality; both are linear in η. The lemma is therefore the paper's fixed-θ LP with the objective $\mathbb{E}[\eta]$ replaced by $\mathbb{E}[h\,\eta]$.

Proposition 6 (Identified set for the regression coefficient)**.** *Let $\mathcal{M} \subseteq \mathbb{R}^{d+2}$ be the set of μ attainable by some feasible η; $\mathcal{M}$ is convex and compact, and is a polytope under score/event (linear) calibration. The sharp identified set for the BLP coefficient is the image*

$$B = \{\Phi(\mu; m_{\text{obs}}) : \mu \in \mathcal{M}\}.$$

Each coordinate of B is computed by the fixed-value method of Section 8. By Frisch–Waugh–Lovell, $\gamma = \mathbb{E}[\tilde{V}\,\eta] \,/\, D(\theta, \mathbb{E}[X^* Z])$, where $\tilde{V}$ is the residual of V on $(1, Z)$ (observed) and $D = \mathbb{E}[\tilde{X}^{*2}] \geq 0$ depends only on the first-stage moments $(\theta, \mathbb{E}[X^* Z])$. Fixing those first-stage moments at any value in their calibrated region makes $\gamma = c$ linear in η, so B is traced by a parametric family of LPs indexed by that low-dimensional region.

Replacing $\mathcal{M}$ by the box of component-wise intervals from Lemma 1 gives valid but generally conservative outer bounds; the joint program is sharp. This is the same distinction drawn for the named-vector LP versus the marginal intersection in Section 10.

Remark 9 (Conditional versus marginal calibration)*.* The bounds are valid under the calibration of Section 4, which constrains η only through its U-marginal and so permits η to vary freely with (Z, V) within report cells. If instead calibration is validated *within* covariate strata—$\mathbb{E}[w(U) \mid X^* = 1, Z = z] \geq a(z)$, the natural design when validation data carry Z—then within-stratum prevalences and outcome cross-moments are bounded separately and the coefficient bounds tighten accordingly. Conditional calibration is to this section what reporter-specific calibration was to Design II: it is where the covariates earn their keep.

Remark 10 (Relation to misclassified-regressor bounds)*.* With U a single noisy report, this is the partial-identification problem for a regression with a misclassified binary regressor (Bollinger, 1996; Mahajan, 2006; Hu, 2008; Molinari, 2008). Our contribution is not a new bound for that problem but the observation that the *same* externally calibrated scores and events that bound prevalence also bound the regression coefficient, through the single channel of the latent cross-moments μ; no independence or instrument is invoked.

A transparent special case. With no covariates and target the mean contrast $\gamma = \mathbb{E}[V \mid X^* = 1] - \mathbb{E}[V \mid X^* = 0]$,

$$\gamma = \frac{\mathbb{E}[V X^*] - \theta\,\mathbb{E}[V]}{\theta(1-\theta)},$$

so γ is pinned down once θ and $\mathbb{E}[V X^*]$ are, each bounded by Lemma 1; the identified set is obtained by ranging the numerator and θ jointly over their calibrated region. If, in addition, V is itself a calibrated high-confidence label, the contrast inherits the sharpness of Section 4.

Inference. Sampling uncertainty enters through the observed law of (U, Z, V) and, for honest calibration, through the validation estimates of the calibration constants. Propagating the bands (15) of Section 8 (or a multiplier bootstrap) through the LPs of Lemma 1 yields a valid outer confidence set for each coordinate of β; moment-inequality methods apply verbatim because all restrictions are linear in η.

12 Practical recommendations

Use calibration as the main identifying content. The maintained assumptions should be calibrated score and event restrictions. Report the source of calibration, preferably lower confidence bounds from validation data on a split disjoint from the main sample.

Treat weak ordering as regularity, not identification. FOSD, MLR, and weak mean ordering are plausible but do not rule out $q_0 = q_1$ and should not be presented as identifying assumptions.

Report sharp sets where they are cheap. For binary scores and events, the linear bounds are already sharp. For graded scores, report the sharp set of Theorem 3; it costs a sort, and the gap from the linear bounds measures the information in the score's shape.

Exploit design symmetries, not independence assumptions. If prompt labels are exchangeable by design, reduce the matrix to the column-pattern histogram N under matched, invariant restrictions (Assumption 1) rather than imposing conditional independence.

Store the richest object possible. For Design I, store all repeated responses even if the count is analyzed. For Design II, store the named vector. For Design III, store the full matrix whenever feasible; under prompt exchangeability, store or construct N; at minimum store the model-count vector. The training-set panel of Section 10 shows what is lost otherwise.

Report coarsening and influence diagnostics. Compare bounds under richer and coarser summaries; report whether results depend on particular models or prompts; quantify losses along $R \to N \to T \to C$ where relevant.

Keep sensitivity assumptions separate. Directional asymmetry, anchored separation, and independence-style restrictions can be useful, but label them as sensitivity analysis unless independently justified.

13 Conclusion

The identifying content of LLM measurement panels depends on the source of replication. Asking one LLM repeated exchangeable questions creates a count experiment; asking several named LLMs once creates a named-reporter experiment; asking several named LLMs repeated questions creates a two-way response matrix. These designs should not be collapsed into a single vote-count model at the outset.

The general theory is simple, and we present it as an adaptation of known mixture and misclassification logic to a setting where dependence among reporters is the rule. Every finite summary yields a two-component mixture for the latent truth; without restrictions, prevalence is completely unidentified, and weak ordering restrictions do not help because they permit equality of the latent components. Useful identification comes from calibrated scores and calibrated events, of which reporter-specific accuracy, threshold rules, relaxed support, unanimity, weighted ensembles, and matrix-agreement events are special cases. The sharpness analysis clarifies exactly what each calibrated object delivers: linear bounds that are sharp for binary scores and for any analysis recording only the score mean, and a rearrangement characterization of the strictly sharper set available from the full score law.

The matrix design adds one further lesson. Even when entries of the matrix are arbitrarily correlated, design symmetries justify lossless reductions: under prompt exchangeability and matched invariant restrictions, the column-pattern histogram preserves all identifying information in the full matrix. Calibration identifies; storage and symmetry determine what can be calibrated without loss.

A Multiple observable groups

Suppose items belong to observable groups $W \in \{1, \dots, G\}$ with known shares ρ_g and group-specific observed histograms $p_g(s) = \mathbb{P}(S = s \mid W = g)$, each group has prevalence $\pi_g = \mathbb{P}(X^* = 1 \mid W = g)$, and the latent class-conditional count distributions are common across groups:

$$p_g(s) = (1 - \pi_g)q_0(s) + \pi_g q_1(s), \qquad g = 1, \dots, G,$$

with target $\theta = \sum_g \rho_g \pi_g$. The common-$q_z$ restriction is an exclusion restriction—groups shift prevalence but not measurement technology—in the spirit of Henry et al. (2014),

and should be used only when measurement errors are plausibly stable across groups. The exact model is bilinear in (π_g, q_z); a conservative LP is obtained from McCormick envelopes for the products $\pi_g q_1(s)$ and $(1-\pi_g)q_0(s)$.

References

Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics*, 73(2):387–399.

Cheng, Z., Wu, X., Yu, J., Han, S., Cai, X.-Q., and Xing, X. (2024). Soft-label integration for robust toxicity classification. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

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

Henry, M., Kitamura, Y., and Salanié, B. (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5(1):123–144.

Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.

Hovsepian, K., Liu, D., and Murugesan, S. (2024). Label with confidence: Effective confidence calibration and ensembles in LLM-powered classification. In *CIKM 2024 Workshop on Generative AI for E-Commerce*.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61.

Linder, F., Leeper, T. J., Haimovich, D., Tax, N., Perini, L., and Vojnovic, M. (2026). Unbiased prevalence estimation with multicalibrated LLMs. *arXiv preprint arXiv:2604.21549*.

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica*, 74(3):631–665.

Manski, C. F. (1999). *Identification Problems in the Social Sciences*. Harvard University Press.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195.