Separating Value Disagreement from Data Uncertainty in Pluralistic Preference Data

Ahmad A. Rushdi [1]

Abstract

Pluralistic preference data entangles two operationally distinct phenomena: genuine value disagreement that should be *preserved* as a multi-modal label, and under-sampled items that need *more annotation*. Standard ensemble-uncertainty estimators conflate the two, treating disagreement as a single signal. We propose a credal disjoint-head model that learns the population-mean preference and a preference-dispersion proxy on separate gradient paths, encouraging a structural separation. Our robust finding is *decorrelation*: on a synthetic generator with closed-form ground truth and a HelpSteer3 disagreement subset, the two estimators stay near-independent where the baseline holds them tightly coupled. Recovering the ground-truth epistemic ranking is a secondary result—clear in the data-rich regime, modest on average—so we foreground decorrelation over recovery. The decomposition supports a candidate per-item routing rule between "collect more annotators" and "preserve disagreement". A pilot held-out annotator simulation shows the rule routes in the predicted direction.

1. Introduction

Pluralistic alignment uses multi-annotator preference data — PRISM (Kirk et al., 2024), HelpSteer3 (Wang et al., 2025), AllenAI MultiPref (Miranda et al., 2024), DICES (Aroyo et al., 2023) — and the field has converged on the position that annotator disagreement is informative signal rather than measurement noise (Aroyo & Welty, 2015; Pavlick & Kwiatkowski, 2019; Sap et al., 2022; Plank, 2022; Davani et al., 2022). But disagreement has *two* sources that demand opposite operator interventions. Figure 1 makes the contrast concrete. An item on which 20 annotators split 10/10 (Item A) is genuine value pluralism that should be *preserved* as a multi-modal target distribution, while an item labeled by only 2 of 20 raters (Item B) is under-sampled and should drive *more annotation*. Aggregating Item A suppresses minority preferences; treating Item B as fixed pluralism wastes the cheapest available intervention.

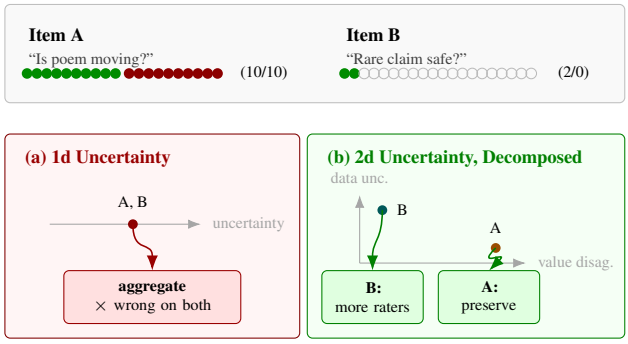

Figure 1. Two items, two reasons annotators disagree. Item A has 20 raters split 10/10 — people genuinely hold different views, and the right move is to *preserve* that spread rather than average it away. Item B has only 2 raters — the right move is to *collect more annotations*. **(a)** Standard methods compress both to one "uncertain" score and action. **(b)** Our model separates reasons for disagreement on a 2-D plane, and items go to different actions.

The standard machinery for telling these regimes apart is the mutual-information decomposition of an ensemble's predictive entropy (Kendall & Gal, 2017; Depeweg et al., 2018; Hüllermeier & Waegeman, 2021; Lakshminarayanan et al., 2017): total predictive uncertainty t splits into aleatoric uncertainty a (the expected per-member entropy) and epistemic uncertainty e (the residual mutual information). Throughout, hats $(\hat{t}, \hat{a}, \hat{e})$ are estimators computed from a finite-size ensemble; unhatted t, a, e are the closed-form ground-truth quantities introduced in §3.1.

Recently, Mucsányi et al. (2024) showed across many ensemble constructions that the resulting $\hat{a}$ and $\hat{e}$ are rank-correlated at 0.8–0.999 and effectively interchangeable on the empirical numbers; the same finding has been reinforced from theoretical (Wimmer et al., 2023; Bickford Smith et al., 2025) and diagnostic (de Jong et al., 2024) angles. Inheriting this pathology silently flattens the very distinction the pluralistic framing requires (Figure 1a): both Items A and B collapse to one "uncertain" scalar and get the same action.

[1]Stanford University, Stanford, CA, USA. Correspondence to: Ahmad A. Rushdi <rushdi@stanford.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

We propose a credal disjoint-head model in which the population-mean head and the annotator-spread head are trained on *non-overlapping* gradient paths, encouraging the two estimators to track separate preference-summary statistics rather than deriving both from a single predictive distribution. The construction is adapted from Credal Concept Bottleneck Models for vision (Mukherjee et al., 2026); the preference setting is a natural fit because the two heads target empirically uncorrelated moment statistics of the per-annotator labels (per-item mean versus per-item entropy).

Our contributions are:[1] (i) we formalize a distinction between disagreement (to be preserved) and uncertainty that should trigger more annotation; (ii) we introduce a disjoint-head preference model whose two estimators target uncorrelated label statistics, so the losses do not implicitly compete; (iii) we provide a synthetic benchmark with closed-form ground truth on which the MI baseline conflates the two axes and ours does not; (iv) we evaluate on a HelpSteer3 disagreement subset and show substantially reduced estimator coupling on real data; (v) we derive a candidate per-item routing rule between aggregation, preservation, reannotation, and refusal (Figure 1b), and verify on a held-out annotator simulation that it routes in the predicted direction.

2. Related Work

The aleatoric/epistemic distinction is canonical in Bayesian deep learning (Kendall & Gal, 2017; Depeweg et al., 2018; Hüllermeier & Waegeman, 2021). Standard practice computes a as the expected per-member entropy of a deep ensemble (Lakshminarayanan et al., 2017) and e as the residual mutual information between the prediction and the choice of ensemble member. This is theoretically motivated only when the ensemble represents a Bayesian posterior over weights — not typically satisfied by deep ensembles — and Mucsányi et al. (2024) catalogue, across many methods and ensemble constructions, that the resulting estimators are rank-correlated at 0.8–0.999 and qualitatively interchangeable. Our synthetic experiment exhibits the same pathology in miniature; the credal construction we propose does not.

A complementary line of work, drawn largely from NLP, argues that annotator disagreement is informative signal rather than noise: Aroyo & Welty (2015) on Crowd Truth, Pavlick & Kwiatkowski (2019) on inherent disagreements in textual inference, Sap et al. (2022) on demographic attitudes biasing toxic-language annotation, Plank (2022) on human label variation as a feature, and Davani et al. (2022) on per-rater modeling beyond majority votes. The position is now operationalized in preference datasets that explicitly preserve per-annotator labels: PRISM (Kirk et al., 2024) (participatory annotation with demographic metadata), HelpSteer3 (Wang et al., 2025) (uniform $K=3$ annotators across $\sim$38 K items), AllenAI MultiPref (Miranda et al., 2024), and DICES (Aroyo et al., 2023) (per-rater safety judgments). Conceptually, the *Value of Disagreement* (Fazelpour & Fleisher, 2025), *Diverging Preferences* (Zhang et al., 2025), and PERSONA (Castricato et al., 2025) argue for preserving rather than aggregating pluralistic disagreement. None of this work makes the uncertainty distinction operationally testable on a single item. The connection between multi-annotator data and the a/e decomposition remained implicit.

The disjoint-gradient construction we use comes from *Credal Concept Bottleneck Models* (Mukherjee et al., 2026), which introduced it for vision concept classification, in the broader credal-deep-learning tradition (Caprio et al., 2024; Hofman et al., 2024). The preference setting is a particularly natural fit: the per-item mean annotator label and the per-item empirical Bernoulli entropy are uncorrelated moment statistics, so the two heads have empirically disjoint targets by construction. Our work also relates to broader work on uncertainty in language generation (Baan et al., 2023; Hou et al., 2024), where similar identifiability concerns arise but on free-form text rather than scalar preference outputs.

A parallel line of work attacks the same conflation by changing the *output* of the reward model rather than its training graph: Distributional Preference Learning (Siththaranjan et al., 2024) fits a distribution of scores per alternative, Probabilistic Uncertain Reward Models (Sun et al., 2025) extract Bradley–Terry reward distributions and explicitly distinguish aleatoric (label inconsistency) from epistemic (out-of-distribution), pairwise-calibrated rewards (Halpern et al., 2025) fit a distribution of reward *functions* matched to per-pair annotator fractions, and PAL (Chen et al., 2025) reduces the parameter cost of personalized rewards. These approaches recover heterogeneity at output; our complementary contribution is a factorization intended to reduce estimator coupling inside the decomposition itself, since the standard MI estimators can remain coupled and operationally interchangeable in practice (Mucsányi et al., 2024).

The downstream use we propose — per-item routing between "preserve disagreement" and "collect more annotators" — connects to active preference acquisition. BALPM (Melo et al., 2024) uses ensemble e plus feature-space entropy to drive label acquisition, and HyPER (Miranda et al., 2025) learns a hybrid router that decides whether each instance should go to humans or LLM judges. Our routing rule differs in that it is read directly off the proposed two-axis decomposition rather than learned end-to-end. A complementary point about the aleatoric floor is operationalized by Abercrombie et al. (2025), who report that annotators disagree with their own prior labels $\sim$25 % of the time — irreducible label noise even under perfect annotator identity.

[1] Code and all experiments: https://github.com/aarushdi/credal-disjoint-heads.

3. Methods

Our methods involve three parts (Figure 2). §3.1 builds a synthetic preference generator with closed-form ground-truth $a(x), e(x)$ — against which any decomposition can be evaluated, since real preference data has no known a, e. §3.2 sets the bar with the standard mutual-information ensemble decomposition — the rank-correlated baseline of Mucsányi et al. (2024). §3.3 introduces our credal disjoint-head model. §4.1 compares the two estimators against the ground truth.

3.1. Synthetic generator with known ground truth

We use a one-dimensional synthetic preference task in which each item is indexed by a scalar feature $x \in [0, 1]$. Concretely, x is a stand-in for "which prompt/response-pair we are looking at" — a single scalar replacing the high-dimensional features (text embeddings, model identities, prompt categories) of a real preference dataset. The benefit of this 1-D parameterization is that we can place a and e as *known* functions of x, evaluate any decomposition method against them on a held-out grid, and visualize the result directly. We return to higher-dimensional features in §4.2 with HelpSteer3 sentence embeddings.

We define two latent functions over x. The *quality difference* $\delta(x)$ controls the population-level direction of preference — a positive δ means the average annotator prefers response A over B at this item. The *annotator value spread* $\sigma_a(x)$ controls how much annotators legitimately disagree at x — a large σ_a means the population holds different values that yield different preferences.

The dataset has two parameters: N, the number of items sampled uniformly on $[0, 1]$, and n_{ann}, the number of independent annotators per item. Each annotator i at item x draws an idiosyncratic offset $z_i \sim \mathcal{N}(0, \sigma_a(x))$ and observes a Bernoulli preference $y_i \sim \text{Bern}\big(\sigma((\delta(x)+z_i)/\tau)\big)$, where σ is the logistic and τ a fixed temperature. We vary N and n_{ann} in §4.1.

We deliberately phase-shift $\delta(x)$ and $\sigma_a(x)$ in x so that the magnitude of the preference signal and the magnitude of irreducible disagreement are not co-located along the input axis. With $p_z(x) = \sigma((\delta(x)+z)/\tau)$ the per-annotator preference probability and $H(\cdot)$ the binary entropy, the closed-form ground-truth quantities are

$$\begin{aligned} t(x) &= H\big(\mathbb{E}_z\, p_z(x)\big), \\ a(x) &= \mathbb{E}_z\big[H(p_z(x))\big], \\ e(x) &= t(x) - a(x). \end{aligned} \tag{1}$$

In our setup a measures the irreducible per-annotator spread under the value-heterogeneity prior, while e measures uncertainty about the population mean given that prior — the residual that shrinks as more annotators are sampled. The classical "aleatoric/epistemic" labels are imperfect for pluralistic preferences (some forms of value heterogeneity are themselves irreducible), but the operational distinction we care about is between disagreement that should be *preserved* and mean-uncertainty that should trigger *more annotation*. The crucial point for the experiment is that a and e are independent design choices: any decomposition that genuinely separates them must recover them independently on a held-out grid.

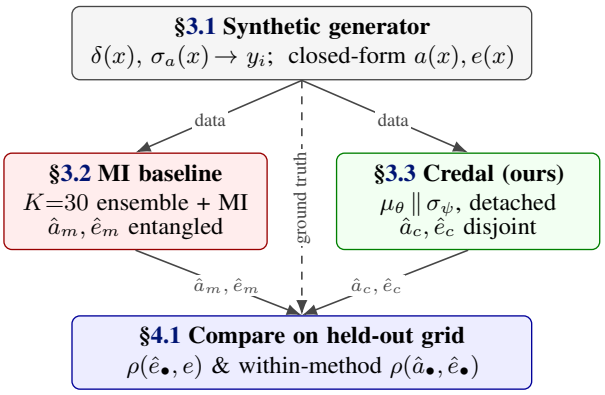

Figure 2. Roadmap. The synthetic generator (§3.1) supplies data to both estimators and the closed-form a, e ground truth that no real preference dataset offers. The MI baseline (§3.2) and our credal model (§3.3) yield estimator pairs $(\hat{a}_m, \hat{e}_m)$ and $(\hat{a}_c, \hat{e}_c)$ respectively; §4.1 compares each against the ground truth.

3.2. Mutual-information baseline

This subsection defines the standard decomposition we benchmark against. We use the textbook ensemble-based mutual-information recipe applied exactly as it would be on real preference data, so that any failure to separate a and e is a property of the recipe rather than of an unfaithful re-implementation. The predictor is an ensemble of K=30 logistic-regression preference classifiers $\{p_k\}_{k=1}^K$ fit on bootstrap resamples of the long-form (item, annotator) data, each using a fixed Fourier feature map $\phi(\cdot)$ of x. Member k produces $p_k(x) \in [0, 1]$, the predicted probability that response A is preferred to B at item x, and we write $\bar{p}(x) = \frac{1}{K}\sum_{k=1}^K p_k(x)$ for the ensemble mean. From the same K predictions, the MI baseline reads off three estimators (subscript m denotes "MI"):

$$\begin{aligned} \hat{t}_m(x) &= H(\bar{p}(x)), \\ \hat{a}_m(x) &= \tfrac{1}{K}\sum_{k=1}^K H(p_k(x)), \\ \hat{e}_m(x) &= \hat{t}_m(x) - \hat{a}_m(x). \end{aligned} \tag{2}$$

Mucsányi et al. (2024) report this exact recipe becomes near-perfectly rank-correlated in practice across vision benchmarks; we predict the same on our preference synthetic.

3.3. Credal disjoint-head model

We propose a network with two parallel output modules — two "heads," each a small MLP that produces one prediction — whose a and e estimators come from *parameter-disjoint* computational paths.[2] The μ-head learns the population-mean preference, the σ-head targets a proxy for irreducible annotator spread, and the two are trained on empirically uncorrelated targets (per-item mean label vs. per-item empirical entropy) so the two losses cannot compete for the same signal. A single forward-only `detach` removes the only path through which one head's gradient could reach the other head — *credal disjointness* by construction. Across a deep ensemble of M random-seed copies, the across-seed variance of the μ-head gives $\hat{e}_c$, and the across-seed mean of the σ-head's entropy estimator gives $\hat{a}_c$. We unpack each piece below.

Concretely, both heads sit on a fixed (non-trainable) Fourier feature map $\phi(\cdot)$ and are otherwise two-layer parameter-disjoint MLPs. The μ-head μ_θ takes $\phi(x)$ and outputs a logit $\hat{\delta}(x)$; the σ-head σ_ψ takes the same $\phi(x)$ and outputs $\log \hat{\sigma}_a(x)$. Disjointness is structural: $\theta \cap \psi = \emptyset$. For item i with n_i annotators and labels $\{y_{ij}\}$, write $\bar{y}_i = n_i^{-1} \sum_j y_{ij}$ for the per-item mean and $H(\bar{y}_i)$ for its binary entropy. The two heads have separate training signals, both derived from the same per-item labels but through uncorrelated moment statistics. The μ-head is trained by binary cross-entropy against $\bar{y}_i$ (never against per-annotator labels), $\mathcal{L}_\mu = \mathrm{BCE}\big(\sigma(\hat{\delta}(x_i)),\, \bar{y}_i\big)$. The σ-head is trained by mean-squared error between its predicted Monte-Carlo entropy:

$$\widehat{H}_\theta(x) = \mathbb{E}_{z\sim\mathcal{N}(0,\,\exp\sigma_\psi(x))}\big[H\big(\sigma(\hat{\delta}(x)+z)\big)\big], \quad (3)$$

and the empirical target $H(\bar{y}_i)$:

$$\mathcal{L}_a = \mathrm{MSE}\big(\widehat{H}_\theta(x_i),\, H(\bar{y}_i)\big). \quad (4)$$

The σ-head thus regresses the empirical per-item entropy $H(\bar{y}_i)$; we treat $\hat{a}_c$ as a *proxy* for a and do not claim population-limit consistency, validating it empirically through the synthetic recovery $\rho(\hat{a}, a)\approx 0.80$ (§4.3). The $\hat{\delta}(x)$ that enters $\widehat{H}_\theta$ in (3) is detached from the computational graph, so $\nabla\mathcal{L}_\mu$ updates only θ, $\nabla\mathcal{L}_a$ updates only ψ, and neither loss can reach across heads. Finally, we wrap the credal model in a deep ensemble of $M{=}8$ random-seed copies $\{(\theta_j, \psi_j)\}_{j=1}^M$, giving per-seed predictions $\hat{\delta}_j(x)$ and Monte-Carlo entropies $\widehat{H}_{\theta_j}(x)$. The credal-method estimators (subscript c) are:

$$\begin{aligned} \hat{a}_c(x) &= \tfrac{1}{M}\sum_{j=1}^{M} \widehat{H}_{\theta_j}(x), \\ \hat{e}_c(x) &= \mathrm{Var}_j\big(\hat{\delta}_j(x)\big), \\ \hat{t}_c(x) &= \hat{a}_c(x) + \hat{e}_c(x). \end{aligned} \quad (5)$$

The structural inversion from (2) is the order: in the MI baseline $\hat{e}_m$ is a residual after $\hat{t}_m$ and $\hat{a}_m$; in the credal model $\hat{a}_c$ and $\hat{e}_c$ are estimated directly on parameter-disjoint paths, and $\hat{t}_c$ is the derived sum.

4. Experiments

We run three main experiments and end with a preliminary routing-rule validation. §4.1 (**E1**) verifies identifiability on the synthetic generator with known ground-truth a, e. §4.2 (**E2**) checks whether the within-method decorrelation property survives on real preference data (HelpSteer3), where ground truth is unavailable. §4.3 (**E3**) ablates the gradient detach to isolate which mechanism — architectural disjointness or training-time gradient routing — drives the result. §4.4 turns the resulting $(\hat{a}, \hat{e})$ pair into a per-item routing rule.

4.1. Synthetic ground-truth recovery

We sweep $n_{\text{ann}} \in \{3, 5, 10\}$ to span the real-data minimum ($n_{\text{ann}}{=}3$ matches HelpSteer3's per-item annotator count) through a data-rich regime, and $N \in \{200, 1000, 5000\}$ items to span data-scarce to data-rich settings (the largest cell yields $N{\cdot}n_{\text{ann}}{=}50{,}000$ long-form rows, comparable to HelpSteer3's training subset). Every combination is evaluated on a fixed held-out grid of 200 query points $x \in [0, 1]$ — dense enough for stable Spearman ρ estimates, small enough that the closed-form a, e are tractable to compute on the grid for evaluation. We report 2 metrics throughout:

Recovery $\rho(\hat{e},\, e)$	**Decorrelation** $\lvert\rho(\hat{a}, \hat{e})\rvert$
rank correlation between an estimator and ground truth on the held-out grid.	within-method rank correlation between estimators on the held-out grid.
Large = good recovery	*Small = clean separation*

Figure 3 reports these two metrics on the headline setting; while Table 1 repeats them across the full sweep. Two patterns are robust across the nine cells of Table 1. The baseline mixes $\hat{a}$ and $\hat{e}$ across the entire sweep, with within-method correlation never below $+0.13$ and reaching $+0.73$ in the data-rich, many-annotator regime ($n_{\text{ann}}{=}10$, $N{=}1000$). The credal model holds the same correlation near zero across the same sweep — the largest absolute value among non-degenerate cells is 0.19 at $n_{\text{ann}}{=}5$, $N{=}5000$ and the

[2] Three quantities are at play—mean preference, value spread a, and model uncertainty e—but only two are *trained* heads (μ, σ); e is read off as the across-seed variance of the μ-head, not a third trained output, which keeps the two training targets uncorrelated.

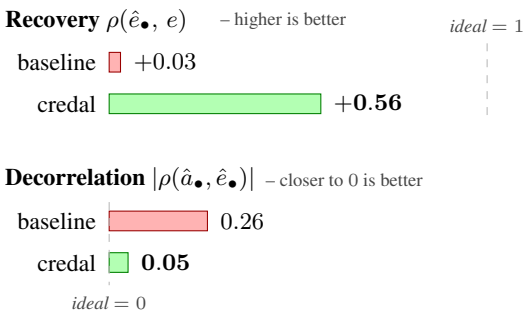

Figure 3. Synthetic experiment E1 at the headline setting (n_{ann}=10, N=5000). The credal model recovers ground-truth e nearly 20× better than the baseline (top) while keeping within-method correlation low (bottom). Both methods recover a similarly well ($\rho(\hat{a}, a)$≈0.80, not shown).

Table 1. Synthetic sweep across n_{ann} and N. **Recovery** (left): credal $\rho(\hat{e}, e)$ matches or beats baseline on most cells, with the headline setting shaded in the bottom row. **Decorrelation** (right): credal $|\rho(\hat{a}, \hat{e})|$ is strictly smaller than baseline in every non-degenerate cell. "–" marks the boundary cell (n_{ann}=3, N=200) where the credal σ-head is numerically degenerate because per-item empirical entropy collapses.

		$\rho(\hat{e}, e)$ (higher better)		$\lvert\rho(\hat{a}, \hat{e})\rvert$ (lower better)	
n_{ann}	N	base	credal	base	credal
3	200	+0.18	–	0.48	–
3	1000	+0.13	+0.08	0.25	0.03
3	5000	−0.21	−0.38	0.13	0.02
5	200	−0.06	+0.15	0.18	0.06
5	1000	−0.29	+0.03	0.20	0.03
5	5000	+0.38	+0.22	0.54	0.19
10	200	+0.10	+0.17	0.28	0.07
10	1000	−0.10	−0.23	0.73	0.07
10	**5000**	**+0.03**	**+0.56**	**0.26**	**0.05**
mean (non-deg.)		+0.02	**+0.07**	0.35	**0.07**

average is 0.04. The contrast is starkest in the headline setting visualized in Figure 3 (n_{ann}=10, N=5000): the credal model decorrelates the two estimators (ρ=+0.05, indistinguishable from zero) and at the same time recovers the ground-truth e ranking (ρ=+0.56); the baseline mixes the two estimators (ρ=+0.26) and barely tracks the ground-truth e at all (ρ=+0.03). At smaller N the credal model is noisier on absolute e magnitudes, as one would expect from any deep-ensemble e estimator with limited data, but its decorrelation property is preserved across the entire sweep. Averaged over the non-degenerate cells the recovery gain is modest ($\bar{\rho}(\hat{e}, e)$=+0.07 vs +0.02) and the credal model loses on three cells; the robust, every-cell property is decorrelation, not recovery.

4.2. Real preference data: HelpSteer3

E1 verified identifiability against closed-form ground truth on the synthetic generator. The natural follow-on is whether

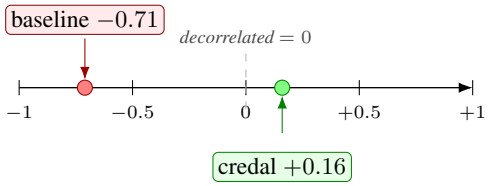

Figure 4. Real-data **decorrelation** on the HelpSteer3 disagreement subset (no ground truth). Position of the within-method correlation $\rho(\hat{a}, \hat{e})$ on the $[-1, +1]$ spectrum: the credal decomposition shifts it from −0.71 (baseline, heavy anti-coupling) to +0.16 (near-independence) on the same 417 held-out items.

the estimator separation survives when we move from a known 1-D generator to real multi-annotator preference data — the operating condition of any deployed alignment pipeline, where ground truth is by definition unavailable. We use HelpSteer3 (Wang et al., 2025), which provides three independent annotator preference scores per (prompt, response-pair) item across approximately 38,000 items; the per-item annotator count n_{ann}=3 matches the lower end of our E1 sweep. With ground truth unavailable, we cannot run the *recovery* diagnostic — only *decorrelation* (the within-method $|\rho(\hat{a}, \hat{e})|$). **Decorrelation alone does not imply correctness**: a model could decorrelate two arbitrary heads without either being meaningful. What this experiment shows is narrower: that the structural property carries from a known 1-D generator to a real, unknown, high-dimensional generator. The synthetic identifiability result (§4.1) is what makes the real-data decorrelation interpretable as separation rather than noise.

The n_{ann}=3 setting introduces a degeneracy: items on which all three annotators agree carry zero empirical entropy and are uninformative about a regardless of method. We therefore restrict evaluation to the *disagreement subset* — 2,082 of 37,241 items where at least one annotator's sign differs, split 1,665/417 train/test. This is the minimal filter that makes the a estimation problem non-degenerate. Each (prompt, response-pair) is encoded with a 768-dimensional `all-mpnet-base-v2` sentence embedding; the bootstrap-ensemble (K=20 logistic-regression heads) and credal-ensemble (M=8 MLPs) are fit on the long-form (item, annotator) data.

Figure 4 answers the motivation's question with a yes, and with a bigger margin than on synthetic: the baseline produces a strongly negative within-method correlation $\rho(\hat{a}, \hat{e})$=−0.71, while the credal model shifts the same number to ρ=+0.16 on the same items. Two qualitative observations are worth flagging. First, the direction of conflation flipped relative to synthetic (+0.31 → −0.71). This is not a contradiction: under the disagreement filter the empirical label-entropy distribution becomes bimodal (most items have entropy 0.637 nats, a few have entropy 0), and the baseline's $\hat{e}$ ends up tracking distance *from* that mode rather

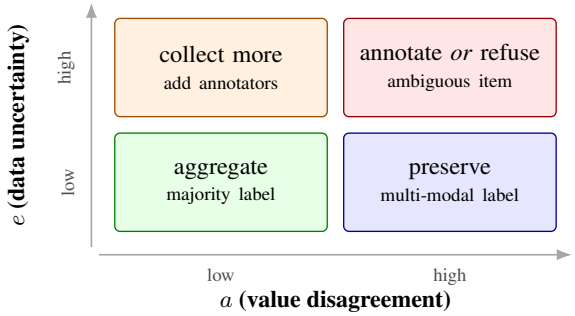

Figure 5. Per-item routing rule under the credal decomposition. Each of the four (a, e) regimes calls for a different deployment-time action. The rule becomes useful only when the two estimators are sufficiently decorrelated and ultimately calibrated; otherwise the off-diagonal routing regimes are unreliable. Under the mutual-information baseline most mass collapses onto a single diagonal axis and the "preserve" and "collect more" quadrants are quasi-empty.

than ground-truth model uncertainty. Second, the credal model's decorrelation property is robust to this bimodal regime, which suggests the disjoint-gradient construction is providing separation in a way that does not depend on the specific shape of the data distribution. Whether the credal $\hat{e}$ on real data tracks any quantity an alignment operator would care about remains the open question (§5); decorrelation alone does not establish that.

4.3. Ablation: detached gradient

Which mechanism drives the result — architectural two-headedness or training-time gradient detachment? We retrain the credal ensemble on the synthetic generator in two configurations: the proposed one (with the detach) and a control without (so gradients from $\mathcal{L}_a$ flow back through both heads). Everything else is identical. **The dominant mechanism is two-headedness on uncorrelated targets, not the detach.** Both configurations decorrelate $\hat{a}$ from $\hat{e}$ ($\rho{=}{+}0.05$ proposed vs. $\rho{=}{-}0.01$ control) and both recover a equally well ($\rho(\hat{a}, a){=}{+}0.80$ in both). The detach contributes a real but modest refinement on e recovery only: $\rho(\hat{e}, e){=}{+}0.56$ proposed vs. $+0.49$ control. We therefore frame the contribution as "factorizing mean preference and preference-dispersion onto separate heads with non-overlapping training targets", with the detach as a refinement, not as the primary mechanism.

4.4. From decomposition to a per-item routing rule

The operational consequence of separating a from e is a per-item routing rule between the two interventions an alignment operator has: collect more annotators, or preserve the existing multi-modal label. Figure 5 shows the four (a, e) regimes and their assigned actions. The diagonal regimes are the easy and hard cases (low-low: aggregate; high-high: ambiguous, annotate or refuse); the two off-diagonal regimes are the load-bearing ones — they distinguish "model uncertainty, fix with more data" (low a, high e) from "irreducible pluralism, preserve the multi-modal label" (high a, low e).

This rule becomes useful only when $\hat{a}$ and $\hat{e}$ are sufficiently decorrelated and, ultimately, calibrated against held-out annotation behavior (the quadrant boundaries in Figure 5 are illustrative median-splits, not calibrated thresholds). Under the MI baseline they move together (Figure 4), the two off-diagonal quadrants are quasi-empty, and the four regimes collapse onto a single diagonal. Under the credal decomposition the same data populates all four quadrants, so a rule of this kind becomes a candidate.

As a preliminary synthetic check on the routing direction, we start from the original $n_{\text{ann}}{=}10$ estimates, reveal an additional 10 held-out annotators on the same items, re-train the credal model at $n_{\text{ann}}{=}20$, and recompute $(\hat{a}_c, \hat{e}_c)$. Median splits on the original estimates define two cohorts. Doubling annotators reduces $\hat{e}_c$ by 91.7% on the high-$\hat{e}$ cohort vs. 67.4% on the high-$\hat{a}$ cohort; $\hat{a}_c$ shifts by $<6\%$ on both. Items flagged "high $\hat{e}$" respond more to annotation than items flagged "high $\hat{a}$" — the direction the rule predicts. This check is synthetic—it reveals additional *simulated* annotators rather than collecting new human labels; validating the rule against real human re-annotation is left to future work (§5).

5. Discussion and limitations

The cleanest reading: decorrelation is necessary but not sufficient for correct identification of a and e. The synthetic experiment anchors the claim — there we verified both decorrelation and that each estimator individually tracks its ground-truth function (the harder claim). On real data we can only verify decorrelation; a sceptical reviewer could argue in principle that the credal model decorrelates while recovering neither, and absent ground truth we have no direct rebuttal. The synthetic identifiability result is what makes the real-data decorrelation interpretable as separation rather than noise.

The synthetic generator is a deliberate 1-D simplification chosen to make identifiability tractable. The conjecture is that the disjoint-gradient construction continues to separate a from e as the input space gets harder, because the construction does not depend on any property of the input representation; we have not validated that in higher-dimensional preference spaces. The HelpSteer3 evaluation restricts to the disagreement subset because at $n_{\text{ann}}{=}3$ binary annotation the empirical entropy collapses on agreement items; a richer annotation regime ($n_{\text{ann}} \geq 5$, ordinal scale) would make the comparison non-degenerate. The held-out annotator simula-

tion in §4.4 is synthetic and validates only the *direction* of the routing rule, not calibrated deployment thresholds. Two comparisons remain for future work: the baseline's $\hat{a}/\hat{e}$ correlation on the *full* HelpSteer3 (not only the disagreement subset), to separate filter effects from distributional ones; and a head-to-head against output-level distributional methods (DPL (Siththaranjan et al., 2024), PURM (Sun et al., 2025)) on the same subset.

6. Conclusion

We argued that pluralistic alignment work conflates two qualitatively different sources of annotator disagreement, and that the conflation is inherited from the standard MI-based decomposition. We proposed a credal disjoint-head model that encourages separate estimation of the two axes — verified on synthetic data with closed-form ground truth and on a real-data subset of HelpSteer3. The decomposition supports a candidate per-item routing rule between "collect more annotators" and "preserve disagreement", and a held-out annotator simulation routes in the rule's predicted direction; full calibration against human annotation is left to future work.

References

Abercrombie, G., Dinkar, T., Cercas Curry, A., Rieser, V., and Hovy, D. Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP*, 2025. arXiv:2301.10684. Annotators disagree with themselves $\sim$25% of the time: a clean operational definition of irreducible aleatoric noise.

Aroyo, L. and Welty, C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. In *AI Magazine*, volume 36, pp. 15–24, 2015.

Aroyo, L., Díaz, M., Homan, C., Prabhakaran, V., Taylor, A., and Wang, D. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. In *NeurIPS Datasets and Benchmarks*, 2023.

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H.-S., Fernández, R., Plank, B., Sennrich, R., Zerva, C., and Aziz, W. Uncertainty in Natural Language Generation: From Theory to Applications. *arXiv preprint*, 2023. arXiv:2307.15703. Surveys uncertainty and disagreement in natural language generation.

Bickford Smith, F., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., and Rainforth, T. Rethinking Aleatoric and Epistemic Uncertainty. In *ICML*, 2025. arXiv:2412.20892. Decision-theoretic critique; popular IT estimators are poor estimators of what they claim to measure.

Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., and Lee, I. Credal Bayesian Deep Learning. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2302.09656. Infinite-ensemble BNN trained via finitely-generated credal sets; explicit AU/EU disentanglement.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In *COLING*, 2025. arXiv:2407.17387.

Chen, D., Chen, Y., Rege, A., and Vinayak, R. K. PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences. In *ICLR*, 2025. arXiv:2406.08469. Modular personalized reward model.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics (TACL)*, 10:92–110, 2022.

de Jong, I. P., Sburlea, A. I., and Valdenegro-Toro, M. How Disentangled are Your Classification Uncertainties? *arXiv preprint*, 2024. arXiv:2408.12175. Orthogonality + consistency as necessary criteria; defines Uncertainty Disentanglement Error (UDE).

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., and Udluft, S. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *ICML*, 2018.

Fazelpour, S. and Fleisher, W. The Value of Disagreement in AI Design, Evaluation, and Alignment. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025. arXiv:2505.07772.

Halpern, D., Micha, E., Procaccia, A. D., and Shapira, I. Pairwise Calibrated Rewards for Pluralistic Alignment. In *NeurIPS*, 2025. arXiv:2506.06298. Distribution over reward functions calibrated to match annotator-preference fractions on every pair.

Hofman, P., Sale, Y., and Hüllermeier, E. Quantifying Aleatoric and Epistemic Uncertainty with Proper Scoring Rules. *arXiv preprint*, 2024. arXiv:2404.12215. Proper-scoring-rule alternative to IT-MI.

Hou, B., Liu, Y., Qian, K., Andreas, J., Chang, S., and Zhang, Y. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling. In *ICML*, 2024. arXiv:2311.08718. Canonical NLP AU/EU decomposition via clarification-augmented ensembling.

Hüllermeier, E. and Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3):457–506, 2021.

Kendall, A. and Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, 2017.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NeurIPS*, 2017.

Melo, L. C., Tigas, P., Houlsby, N., Foster, A., and Rainforth, T. Deep Bayesian Active Learning for Preference Modeling in Large Language Models. In *NeurIPS*, 2024. Uses epistemic uncertainty plus feature-space entropy to acquire preference labels.

Miranda, L. J. V., Wang, Y., Elazar, Y., Kumar, S., Pyatkin, V., Brahman, F., Smith, N. A., Hajishirzi, H., and Dasigi, P. MultiPref: Multi-perspective preference annotations with expert and normal workers. Hugging Face Datasets, `allenai/multipref`, 2024. Released alongside the HyPER routing study; 2 normal + 2 expert worker annotations per item, 10,461 items.

Miranda, L. J. V., Wang, Y., Elazar, Y., Kuznia, K., Gulati, A., Hofmann, V., Suhr, A., Smith, N. A., Hajishirzi, H., and Dasigi, P. Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback. In *ACL*, 2025. arXiv:2410.19133. Trains a router on MultiPref to decide whether each instance should go to humans or LLM-judges.

Mucsányi, B., Kirchhof, M., and Oh, S. J. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In *NeurIPS*, 2024. arXiv:2402.19460.

Mukherjee, T., Kloft, M., Marquis, P., and Bouraoui, Z. Credal Concept Bottleneck Models: Structural Separation of Epistemic and Aleatoric Uncertainty. *arXiv preprint*, 2026. arXiv:2602.11219.

Pavlick, E. and Kwiatkowski, T. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7:677–694, 2019.

Plank, B. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling, and Evaluation. In *EMNLP*, 2022.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection. In *NAACL*, 2022.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *ICLR*, 2024. arXiv:2312.08358. RLHF implicitly aggregates via Borda count; DPL fits a distribution of scores per alternative.

Sun, W., Cheng, X., Yu, X., Xu, H., Yang, Z., He, S., Zhao, J., and Liu, K. Probabilistic Uncertain Reward Model. *arXiv preprint*, 2025. arXiv:2503.22480. Reward distributions from pairwise preferences; distinguishes label inconsistency from out-of-distribution uncertainty.

Wang, Z., Zeng, J., Delalleau, O., Shin, H.-C., Soares, F., Bukharin, A., Evans, E., Dong, Y., and Kuchaiev, O. HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. *arXiv preprint*, 2025. arXiv:2505.11475. 3 annotators per item, 38,459 train + 2,017 validation.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? In *UAI*, 2023. arXiv:2209.03302. Foundational theoretical critique of IT-decomposition incoherencies.

Zhang, M. J. Q., Wang, Z., Hwang, J. D., Dong, Y., Delalleau, O., Choi, Y., Choi, E., Ren, X., and Pyatkin, V. Diverging Preferences: When Do Annotators Disagree and Do Models Know? In *ICML*, 2025. arXiv:2410.14632.