Beyond Consensus: Perspectivist Modeling and Evaluation of Annotator Disagreement in NLP

Anonymous ACL submission

Abstract

Annotator disagreement is widespread in NLP, particularly for subjective and ambiguous tasks such as toxicity detection and stance analysis. While early approaches treated disagreement as noise to be removed, recent work increasingly models it as a meaningful signal reflecting variation in interpretation and perspective. This survey provides a unified view of disagreement-aware NLP methods. We first present a domain-agnostic taxonomy of the sources of disagreement spanning data, task, and annotator factors. We then synthesize modeling approaches using a common framework defined by prediction targets and pooling structure, highlighting a shift from consensus learning toward explicitly modeling disagreement, and toward capturing structured relationships among annotators. We review evaluation metrics for both predictive performance and annotator behavior, and noting that most fairness evaluations remain descriptive rather than normative. We conclude by identifying open challenges and future directions, including integrating multiple sources of variation, developing disagreement-aware interpretability frameworks, and grappling with the practical tradeoffs of perspectivist modeling.

1 Introduction

NLP applications largely rely on supervised learning, which depends on annotated data. Annotation is often operationalized through majority voting, an assumption that can be problematic—particularly for complex tasks where true experts may be outnumbered, or for inherently ambiguous tasks that admit multiple valid interpretations. Majority aggregation can also obscure minoritized perspectives, leading to biased models and representational harm (Blodgett et al., 2020; Gordon et al., 2021). This "single ground truth" assumption has been increasingly challenged, especially with the rise of subjective NLP tasks such as toxic language detection and quality estimation. Early critiques appear in Aroyo and Welty (2015), which questions the existence of a unique truth in crowdsourced annotation. This shift aligns with broader efforts to make NLP systems more inclusive, as consensus-based labels often disadvantage minority viewpoints (Blodgett et al., 2020; Gordon et al., 2021). Building on this, Cabitza et al. (2023) formalized *perspectivism* in NLP, advocating for models that integrate diverse human viewpoints rather than collapsing them into gold standards.

We formalize a taxonomy of disagreement by synthesizing prior work into three sources: data, task, and annotator factors. We trace the evolution of NLP approaches for learning from disagreement, from latent-truth models to multi-annotator, embedding-based methods. To unify these developments, we introduce a synthesis table highlighting key trends, including the shift from treating disagreement as noise to modeling it as a prediction target, and the growing emphasis on structured relationships among annotators. We also survey evaluation practices and fairness considerations. Finally, we outline open challenges and future directions, including jointly modeling multiple sources of variation and navigating the practical tradeoffs of perspectivist approaches. Our contributions are:

1. **A unified taxonomy of disagreement** across data-, task-, and annotator-driven sources.

2. **A synthesis of disagreement-aware methods**, mapping approaches to disagreement sources, prediction targets, and pooling structures.

Two recent surveys are closely related to our work but differ in scope and emphasis. Frenda et al. (2024) provides a broad, conceptual overview of perspectivist NLP, focusing on definitions, dataset practices, and sociotechnical motivations. We build on their framing while shifting toward a more model-centric and technical analysis. Uma et al.

(2021) survey disagreement-aware learning across computer vision and NLP, emphasizing empirical training and evaluation strategies. Our survey complements this work by centering NLP-specific sources of subjectivity and by jointly examining models and evaluation within a unified framework. We survey over 120 prior works spanning disagreement sources, modeling methods, and evaluation paradigms. We discuss our survey scope and selection criteria in A.1.

2 Sources of Disagreement

Prior work has analyzed sources of annotator disagreement across tasks and domains. Basile et al. (2021) identify three broad contributors: individual differences (partly linked to sociodemographics), stimulus characteristics (e.g., linguistic or task ambiguity), and contextual inconsistency in human behavior. Sandri et al. (2023) propose a finer taxonomy (sloppy annotations, ambiguity, missing information, and subjectivity) and show that in offensive language detection, subjectivity driven by personal bias and task design dominates disagreement. In legal NLP, Xu et al. (2024b) introduce a taxonomy highlighting genuine ambiguity, narrative uncertainty, and annotation context. Overall, existing taxonomies are largely domain-specific, emphasizing different facets of disagreement depending on task formulation and application context.

2.1 Taxonomy of disagreement

We propose a unifying, domain-agnostic taxonomy (Fig. 1) that synthesizes these views into three overarching sources of disagreement: data factors (e.g., ambiguity and data quality), task factors (e.g., formulation and interface design), and annotator factors (e.g., demographics, preferences, and errors). Importantly, different sources of disagreement can interact with one another. Unclear task formulation can lead to inconsistent annotator behavior. Linguistic ambiguity can give rise to disagreement that varies with annotators' individual and group identities, leading to systematically different interpretations of the same text.

2.2 Data Factors

Prior work identifies three data-driven sources of disagreement: data quality issues, linguistic ambiguity, and epistemic uncertainty—cases with no single ground truth, such as subjective or culturally situated interpretations (Poesio, 2020; Uma et al., 2021; Pavlick and Kwiatkowski, 2019). Missing context or noisy inputs exacerbate inconsistency, especially for short text (Plank et al., 2014). More fundamentally, language is inherently ambiguous, and in many subjective tasks disagreement reflects irreducible multiplicity rather than annotation error (Pavlick and Kwiatkowski, 2019; Uma et al., 2021).

2.3 Annotator Factors

Individual identities. Prior work documents stable, annotator-specific biases that introduce variance into training data and directly affect model performance (Otterbacher, 2018; Larimore et al., 2021; Pavlick and Kwiatkowski, 2019; Geva et al., 2019). Beyond general disagreement, psychological and moral traits systematically correlate with annotation behavior. Personality dimensions such as agreeableness and conscientiousness are associated with evaluative leniency (Mieleszczenko-Kowszewicz et al., 2023), while moral values exert stronger influence on judgments of offensiveness than geographic or cultural background (Davani et al., 2024). Together, these findings show that annotators bring stable individual preferences that shape how they interpret and label language.

Group identities. Group-level attributes such as gender, race, age, and political orientation influence annotation behavior, particularly in socially grounded tasks. Gender-linked variation has been observed in hate speech detection (Wojatzki et al., 2018) and even in syntactic tasks (Garimella et al., 2019), aligning with sociolinguistic evidence of gendered language use (Mondorf, 2002). In subjective domains such as toxicity or sentiment, demographic shifts can substantially affect model performance (Ding et al., 2022). Qualitative work further shows that lived experience shapes interpretation: community insiders label gang-related language differently from academic annotators (Patton et al., 2019), and political or racial attitudes correlate with toxicity judgments, including treatment of African American English (Sap et al., 2022; Sang and Stanton, 2021; Luo et al., 2021). In these settings, disagreement often reflects social perspective rather than noise (Chulvi et al., 2023). However, identity effects are task-dependent. For less socially salient tasks (e.g., word similarity, sentiment, NLI), demographic attributes do not consistently explain annotation differences (Biester et al., 2022), and explicitly modeling sociodemographics may yield limited gains (Orlikowski et al., 2023). These findings caution against ecological fallacies (Robinson,

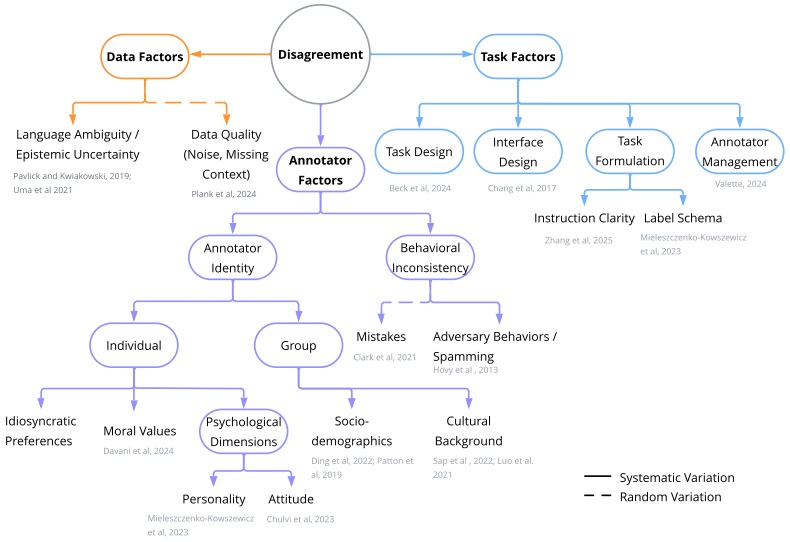

Figure 1: Taxonomy of sources of annotator disagreement, with key citations for each sub-category. For each source of disagreement, we denote systematic variation using solid lines, and random variation using dotted.

1950): group-level identity does not uniformly predict individual annotation behavior.

Annotator behavioral inconsistencies. Abercrombie et al. (2025) found that annotators give inconsistent responses around 25% of the time across four different NLP tasks. Clark et al. (2021) also show that untrained annotators identify GPT-3- versus human-written text only at chance levels. Annotators may also produce low-quality or strategically chosen labels to maximize pay (Hovy et al., 2013).

2.4 Task Factors

Annotation task design—its formulation, structure, and presentation—strongly shapes disagreement. Ambiguous or underspecified tasks permit multiple valid interpretations, leading to systematic divergence; in LLM-as-Judge settings, ambiguous prompts elicit different interpretive strategies (Zhang et al., 2025b). Interface design also affects agreement: unclear guidelines increase inconsistency, while alternative interfaces can surface ambiguity rather than suppress it (Chang et al., 2017). Annotation schemas and scales further modulate disagreement: finer-grained ratings amplify variability relative to binarized labels (Mieleszczenko-Kowszewicz et al., 2023), and different schemas (e.g., scalar ratings vs. pairwise preferences) capture related but distinct judgments, particularly in RLHF (Dsouza and Kovatchev, 2025). Presentation and labor conditions likewise matter: order effects, fatigue, and satisficing shift responses over time (Strack, 1992; Krosnick et al., 1996; Galesic and Bosnjak, 2009; Beck et al., 2024).Fine-grained schemes may inflate disagreement due to degraded consistency under poor working conditions (Valette, 2024; Yang et al., 2023).

3 Learning from Disagreement

We discuss three families of disagreement modeling, tracing NLP's shift from treating disagreement as noise to modeling it as structured variation.

3.1 Latent Truth and Annotator Reliability Modeling

Models in this family attribute disagreement to annotator reliability, bias, or task difficulty, with the goal of inferring latent true labels while estimating annotator behavior. Foundational work by Dawid and Skene (1979) introduced an EM-based framework to jointly infer latent labels and annotator error rates from noisy annotations, underpinning many later Bayesian and discriminative models. Subsequent work incorporates task characteristics, explicitly modeling difficulty (Whitehill et al., 2009; Ma et al., 2015) or latent topics that align with annotator expertise (Fan et al., 2015; Ma et al., 2015; Zhao et al., 2015; Welinder et al., 2010). Parallel work has focused on annotator modeling, representing competence as scalar accuracies (Demartini et al., 2012; Karger et al., 2011; Liu et al., 2012), confusion matrices (Dawid and Skene, 1979; Raykar et al., 2010; Liu et al., 2012), or bias–variance decompositions (Welinder et al., 2010; Raykar et al., 2010). Confusion-matrix-

based models are more expressive and typically outperform simpler formulations (Zheng et al., 2017). Formally, latent truth models posit a single hidden label z_i for each item and represent annotator disagreement as noise around this truth. These approaches model annotations as

$$p(y_{ij}|z_i, a_j) \tag{1}$$

a_j denotes annotator-specific parameters (such as a confusion matrix), and marginalize over z_i during inference. The objective function (Dawid and Skene, 1979; Hovy et al., 2013) maximizes the marginal likelihood of the observed annotations by marginalizing over a single true label per item.

$$\max_{\{a_j\}} \sum_i \log \sum_{z_i} p(z_i) \prod_j p(y_{ij} \mid z_i, a_j) \tag{2}$$

Disagreement reflects annotator unreliability rather than item ambiguity.

Later work integrates these ideas into learning architectures. Neural extensions such as CrowdLayer (Rodrigues and Pereira, 2018) and label-transfer approaches (Tanno et al., 2019) replace explicit confusion matrices with differentiable components, while other models introduce worker weighting or structured noise modeling (Gao et al., 2022; Cao et al., 2023; Wei et al., 2022; Chu et al., 2021). Paun et al. (2018a) show that partial pooling—i.e., drawing annotator parameters from a shared population distribution—best balances expressivity and generalization. Despite their strengths, latent truth models can struggle with scalability and estimation under sparse labeling or weakly differentiated answers. To address this, CROWDLAB (Goh et al., 2023) combines annotator statistics with task-level features by treating classifier predictions as an additional annotator, improving robustness without iterative inference. Recent work further incorporates task features and demographic structure: Simpson et al. (2015) jointly models text, annotator bias, and latent truth, while NUTMEG (Ivey et al., 2025) extends truth inference to demographic subpopulations, estimating competence and predicting labels per group rather than collapsing to a single consensus (Hovy et al., 2013; Paun et al., 2018b).

3.2 Task-based Annotator Models

This model family treats each annotator as a distinct task, modeling labeling behavior directly rather than collapsing annotations into a single truth. This family reframes disagreement from noise to a structured signal. Given an input x_i, a shared encoder produces a representation $h_i = f(x_i)$, and each annotator j defines a task-specific predictor

$$p^{(j)}(y \mid x_i) \tag{3}$$

Model parameters are learned by minimizing a supervised loss over annotator–item pairs,

$$\min_\theta \sum_i \sum_{j \in \mathcal{A}_i} \ell\Big(y_{ij},\ p_\theta^{(j)}(y \mid x_i)\Big) \tag{4}$$

θ is all model parameters (shared encoders and annotator-specific heads), $\mathcal{A}_i$ is set of annotators who labeled i, and $\ell(\cdot, \cdot)$ is typically cross-entropy for classification or squared error for regression. Disagreement is captured implicitly through divergence across annotator-specific predictions.

Early work by Cohn and Specia (2013) models annotators as correlated tasks in a Gaussian Process framework, using an inter-annotator covariance matrix to control pooling. This nonparametric analogue to hierarchical Bayesian models identifies annotator similarity and outliers, with partial pooling performing best (Paun et al., 2018b). Neural extensions implement this idea via shared backbones and annotator-specific heads, including CrowdLayer (Rodrigues and Pereira, 2018) and related architectures (Guan et al., 2018). Fornaciari et al. (2021) further reinterpret disagreement as a regularization signal to reduce overconfidence on ambiguous inputs, learning from both the gold labels and the distribution over multiple annotators (which they treat as soft label distributions in a single auxiliary task). More recent work models richer annotator structure. Mixture-of-experts approaches capture feature-level heterogeneity (Han et al., 2025), while multi-head Transformers preserve uncertainty aligned with empirical disagreement (Mostafazadeh Davani et al., 2022). Group-aware and loss-based extensions incorporate demographic structure or explicit tradeoffs between denoising and minority perspectives (Fleisig et al., 2023; Jinadu and Ding, 2024). Collectively, task-based models capture structured perspective variation beyond latent-truth formulations.

3.3 Embedding-based Annotator Models

The third family of embedding-based models departs from latent-truth paradigms by treating disagreement as systematic variation in annotator behavior rather than noise around a hidden true label. Instead of assigning each annotator a separate prediction head, these approaches encode annotator

differences in a shared latent space, enabling scalability to thousands of annotators with sparse labels, an important limitation of task-based methods. Embedding-based models parameterize annotator-level predictions as

$$p(y_{ij} \mid h_i, e_j) \tag{5}$$

$h_i = f(x_i)$ is an embedding of the input item and $e_j = g(a_j)$ is an embedding of the annotator j. The loss function optimizes standard supervised losses over annotator–item pairs, where θ is all model parameters (shared encoders and annotator-specific heads), $\mathcal{A}_i$ is the set of annotators who labeled i, and $\ell(\cdot, \cdot)$ is typically instantiated as cross-entropy for classification or squared error for regression.

$$\min_{\theta} \sum_{i} \sum_{j \in \mathcal{A}_i} \ell(y_{ij},\ p_\theta(y \mid h_i, e_j)) \tag{6}$$

Disagreement is formalized as the interaction between item and annotator embeddings.

Early work introduced annotator embeddings to capture individual biases in subjective NLP tasks (Kocoń et al., 2021). AART (Mokhberian et al., 2024) extends this idea to large-scale crowdsourcing with regularization for robustness and fairness. Subsequent models enrich the framework by incorporating annotation embeddings (Deng et al., 2023) or demographic structure. Jury Learning (Gordon et al., 2022) combines content, annotator, and demographic embeddings to predict individual votes and compose configurable juries, while DEM-MoE (Xu et al., 2025) explicitly models demographic structure via a mixture-of-experts to capture intersectional variation. Recent work further improves scalability by replacing stored embeddings with hypernetwork-generated, annotator-specific LoRA adapters (Ignatev et al., 2025). Embedding-based models show a progression in the granularity of disagreement they represent: from individual annotators (Kocoń et al., 2021), to demographic groups (Gordon et al., 2022), to population-level distributions (Weerasooriya et al., 2023). This trajectory points toward an important future direction: estimating subgroup-specific disagreement rather than averaging across populations. Building on prior work (Sap et al., 2022; Lakkaraju et al., 2015), future work could focus on inferring how social subgroups interpret the same item, enabling perspective-aware predictions that surface when disagreement reflects broader social divides.

Direct disagreement modeling. A subset of embedding-based models *predicts* disagreement directly rather than treating it as noise. These approaches vary in how disagreement is represented—implicitly via soft supervision, as a scalar score, or explicitly as a label distribution—but share the goal of modeling disagreement as the prediction target. Embedding-based architectures are well-suited to this setting because they enable generalization across annotators, conditioning on demographic attributes, and marginalization over annotator populations. Implicit approaches predict aggregate disagreement without explicitly modeling a distribution. For example, Xu et al. (2024a) predict empirical label proportions from text using cross-entropy against soft labels, while Wan et al. (2023) conditions predictions on annotator demographics and regresses a scalar disagreement score. In contrast, distributional modeling treats disagreement itself as the target, enabling prediction even when not all annotators are observed by directly parameterizing an item-level label distribution

$$p(y \mid x_i) \tag{7}$$

y is the expected distribution of annotators that would be produced by the population, for instance x_i. Distributional models typically optimize a composite objective (Parappan and Henao, 2025; Weerasooriya et al., 2023), combining supervised alignment with individual annotations and a divergence-based term (e.g., KL or JSD) that aligns predicted and empirical item-level distributions:

$$\begin{aligned} \min_{\theta} & \underbrace{\sum_{i} \sum_{j \in \mathcal{A}_i} \ell(y_{ij},\ p_\theta(y \mid x_i, a_j))}_{\text{annotator-level alignment}} \\ & + \lambda \underbrace{\sum_{i} \mathcal{D}(\tilde{p}_i(y) \parallel p_\theta(y \mid x_i))}_{\text{distributional alignment}} \end{aligned} \tag{8}$$

DisCo (Weerasooriya et al., 2023) models disagreement at the response, item, and population levels, aggregating over annotators at inference to estimate population-level disagreement. The Learning Subjective Label Distribution framework (Parappan and Henao, 2025) extends this by conditioning distributions on sociodemographic attributes and semantic perspectives, enabling subgroup-specific disagreement estimates. Collectively, these approaches move toward unified representations of individual-, group-, and population-level variation.

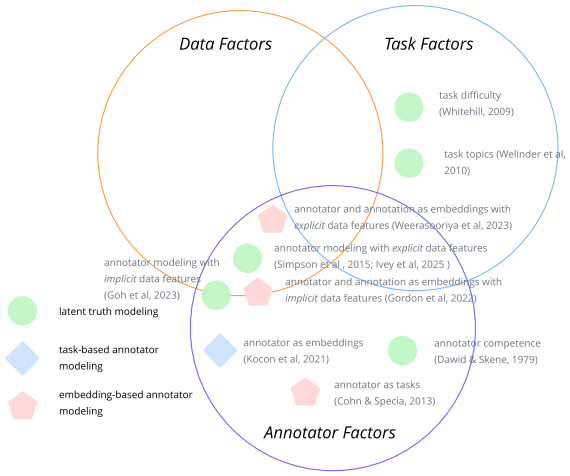

Figure 2: The distribution of the three main methods across our identified taxonomy of sources of disagreement. The lack of work modeling data and task factors point to future directions.

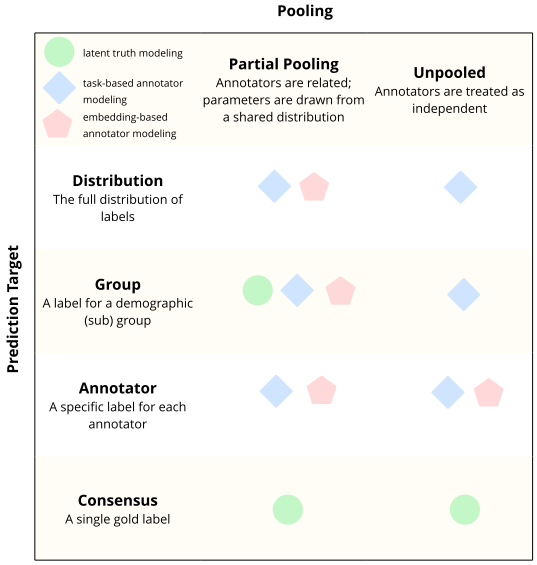

Figure 3: Prediction targets and pooling assumptions across methods for learning from annotator disagreement. The table maps existing approaches by what they predict (consensus labels, individual annotator responses, group-level outputs, or full disagreement distributions) and how they pool information across annotators. Fig. 4 is a more detailed table with representative work.

4 Mapping Models to Sources of Disagreement

We map the three model families onto our taxonomy of disagreement sources in Figure 2. Latent truth models span multiple regions, capturing task factors (e.g., difficulty), annotator factors (e.g., competence), and approaches that condition annotator behavior on the input. Within the intersection of annotator and data modeling, we distinguish between implicit and explicit data features. Implicit approaches incorporate the input via shared encoders or embeddings without explicitly modeling item ambiguity as a source of disagreement (Goh et al., 2023; Raykar et al., 2010; Rodrigues and Pereira, 2018; Gordon et al., 2022). For example, CrowdLayer (Rodrigues and Pereira, 2018) and embedding-based models (Gordon et al., 2022) condition predictions on text but treat it primarily as contextual information. In contrast, explicit data modeling treats properties of the input itself as a structured source of annotator variation (Simpson et al., 2015; Ivey et al., 2025). For instance, Ivey et al. (2025) condition annotator competence on item features to capture systematic subgroup responses. This category also includes distributional approaches that directly predict item-level disagreement distributions, explicitly representing data ambiguity (Weerasooriya et al., 2023; Parappan and Henao, 2025). Notably, the relative scarcity of work modeling data factors alone—or jointly integrating data with task and annotator factors—highlights promising directions for future research.

5 Prediction Targets and Pooling Structures

We introduce a synthesis table (Figure 3) organized along two dimensions: prediction target and pooling structure. Prediction targets include consensus (a single label per item), annotator (one prediction per annotator), group (one prediction per annotator group), and distribution (the full label distribution across annotators). Pooling (Paun et al., 2018c) is generalized across three model classes. Group pooling forces all annotators to share a single labeling function. Unpooled models treat annotators independently. Partial pooling encodes structured relationships among annotators or between annotators and the population (or subpopulations), though it is instantiated differently across paradigms. In Bayesian latent-truth models, annotator reliabilities are drawn from a shared population prior. Gaussian process multi-task models (Cohn and Specia, 2013) treat annotators as correlated tasks, while neural multi-task models (Rodrigues and Pereira, 2018) use shared backbones with annotator-specific heads. Embedding-based approaches (Mokhberian et al., 2024) instead represent annotators as points in a latent population space. Notably, none of the

surveyed disagreement-aware models employ full group pooling. While group pooling corresponds to majority vote or fixed aggregation, once annotator behavior is modeled, methods necessarily adopt unpooled or partially pooled structures. We highlight these temporal trends:

The field is gradually formalizing disagreement as an object of prediction, not a source of noise. The field of modeling annotation disagreement has moved from consensus prediction to individual annotator modeling, to group-aware modeling, and recently to population or subgroup distribution modeling. There is a consistent upward shift on the prediction target axis of the table over time.

Across the pooling structures, the field has moved from unpooled to partial pooling. The most successful models use partial pooling, such as hierarchical priors (Paun et al., 2018c), inter-annotator kernels (Cohn and Specia, 2013), demographic-informed MoE (Xu et al., 2025), and subgroup-aware distributions (Parappan and Henao, 2025), moving toward improved generalization and fairness. The partial pooling of annotator or group is the most expressive and successful. Most existing methods fall into these two cells. This combination would give generalization to new annotators, the ability to reason about groups, and the ability to scale with sparse labels.

Few works actually model disagreement distributions, but there is a convergence toward distribution-level modeling in newer architectures (Fornaciari et al. (2021) as a task-based model, Parappan and Henao (2025); Weerasooriya et al. (2023) as embedding-based models). Predicting how much people disagree is becoming as important as predicting what label they choose.

6 Evaluation Metrics

Evaluation metrics for modeling disagreement broadly fall into two classes: those that compare the predictions to the true annotations, and those that evaluate annotator behavior, aiming to recover latent structure such as inter-annotator relationships. Hard metrics (accuracy, precision, recall) assume a single gold label—problematic for subjective or ambiguous tasks (Poesio and Artstein, 2005; Plank et al., 2014)—motivating a shift toward probabilistic and soft metrics. Aggregation models such as Dawid–Skene (Dawid and Skene, 1979) and MACE (Hovy et al., 2013) use likelihood or accuracy against inferred latent truth, while latent-truth and reliability models adopt ROC–AUC, negative log probability, predictive density, and entropy-based measures (Raykar et al., 2010; Paun et al., 2018c; Simpson et al., 2015; Ivey et al., 2025). As models predict heterogeneous judgments, evaluation extends to per-annotator metrics (MAE, RMSE, macro F1) (Cohn and Specia, 2013; Mostafazadeh Davani et al., 2022) and group-level/distributional divergences (KL, JSD) (Mokhberian et al., 2024; Gordon et al., 2022; Weerasooriya et al., 2023). However, recent work (Rizzi et al., 2024) has shown that Cross Entropy (and to a lesser extent divergence-based measures) may violate desirable properties like symmetry or fair penalization, potentially obscuring meaningful differences between models that aim to reproduce disagreement distributions.

A complementary set of metrics evaluates annotator quality directly. Likelihood-based measures and annotator–model correlations assess inferred behavior (Passonneau and Carpenter, 2014; Paun et al., 2018c), while agreement metrics like Krippendorff's α and Cohen's κ quantify consistency but conflate ambiguity with unreliability (Krippendorff, 1980; Viera and Garrett, 2005). CrowdTruth (Inel et al., 2014) treats disagreement as signal via worker-, item-, and label-level metrics, later extended to model their interactions (Dumitrache et al., 2018). Recent work targets latent structure more directly, using intra-annotator consistency (Gordon et al., 2021) and relational metrics (DIC, BAE) to test whether models preserve agreement geometry (Zhang et al., 2025a). Overall, evaluation increasingly emphasizes models' preservation of disagreement structure, not merely match labels.

Evaluating Fairness. Fairness becomes relevant when disagreement is systematic and socially structured, reflecting genuine differences in perspective tied to demographics or lived experience (Aroyo and Welty, 2013; Pavlick and Kwiatkowski, 2019). Collapsing annotations into a single consensus can erase minority viewpoints, rendering pooling a normative—rather than purely technical—choice. Although fairness is increasingly invoked in disagreement-aware modeling, most evaluations remain descriptive (Binns, 2018). Disaggregated and group-level metrics reveal disparities (Gordon et al., 2022; Xu et al., 2025), but do not specify what outcomes should count as fair. Diagnostic frameworks such as AART (Mokhberian et al., 2024) and PERSEVAL (Lo et al., 2025) report parity gaps without explicit normative ground-

ing. This gap motivates adapting fairness-in-ML frameworks—such as statistical parity, equalized opportunity, and subgroup fairness (Caton and Haas, 2024)—to disagreement-aware modeling, moving toward explicit normative reasoning about what kind of fairness models ought to achieve.

7 Challenges and Future Work

LLM-based simulation of annotator variation. Persona prompting and demographic conditioning with LLMs are increasingly used to simulate annotator variation, but remain methodologically limited. These approaches are largely unpooled (e.g., personas are simulated independently), highly sensitive to prompts and model choice, and risk amplifying stereotypes (Lee et al., 2023; Durmus et al., 2023; Santurkar et al., 2023; Blodgett et al., 2020). Empirically, persona variables explain only a small fraction of variance (Hu and Collier, 2024), and fine-tuning with demographic metadata appears to rely more on annotator-specific signals than generalizable group structure (Orlikowski et al., 2025). More broadly, LLM judgments compress human disagreement and rely on opaque priors, raising concerns about reliability and epistemic validity of replacing human annotation labor (Durmus et al., 2023; Cazzaniga et al., 2024).

Modeling task-induced disagreement. While annotator- and data-level variation are increasingly modeled, task-level sources of disagreement remain underexplored. Instruction phrasing, label schema design, presentation order, and interface choices can systematically shape disagreement, yet are rarely represented explicitly. Prior work shows that task design strongly influences annotation outcomes (Zhang et al., 2025b; Dsouza and Kovatchev, 2025), but most models incorporate task information only indirectly. Future work could embed task schemas directly, enabling models to capture how task framing interacts with annotators and data.

Integrating task, annotator, and data variation. Most existing methods focus primarily on annotator bias as the source of disagreement. A few latent-truth models and embedding models jointly consider annotator and task factors, but task-based approaches typically do not. Developing models that integrate all three sources, task, annotator, and data, remains an open challenge, partly constrained by data availability.

Scarcity of detailed annotation data. Modeling nuanced disagreement, particularly for minoritized groups, is constrained by sparse annotator metadata and noisy labels, highlighting the need for richer, disaggregated annotation data.

Interpretability of disagreement-aware models. Explainability remains underdeveloped for disagreement-aware NLP models. Most methods for model interpretability target single-label predictions and do not explain why annotators or groups disagree (Molnar, 2020). While recent multi-annotator models offer feature- or attention-based explanations, these are rarely validated against real annotator behavior. This gap is especially salient for demographic-aware and mixture-of-experts models, where expert specialization may reflect spurious correlations. Emerging metrics such as Behavior Alignment Explainability (Zhang et al., 2025a) point toward more principled evaluation, but robust disagreement-aware interpretability remains an open problem.

Generalization and practical tradeoffs. Annotator disagreement is highly dataset- and domain-specific, and evidence on the benefits of demographic conditioning is mixed (Gordon et al., 2022; Orlikowski et al., 2023). Models risk overfitting to dataset-specific noise, suggesting a need for cross-dataset and cross-domain training. Perspectivist models introduce tradeoffs: they require richer annotations and greater computation, complicate filtering decisions, and raise normative questions about whose perspectives should be preserved. As a result, adopting disagreement-aware modeling is not merely a technical choice, but a value-laden one balancing fairness, interpretability, efficiency, and annotation labor conditions (Valette, 2024).

8 Conclusion

We present a unified view of disagreement-aware modeling in NLP, focusing on disagreement as a meaningful signal rather than noise. We introduce a taxonomy of data-, task-, and annotator-driven sources and trace the shift from aggregation to multi-annotator and persona-conditioned models. Our synthesis highlights key gaps, including limited distributional modeling and largely descriptive fairness evaluation. We also highlight future directions, such as integrating multiple sources of disagreement and developing interpretable, normatively grounded evaluation frameworks.

9 Limitations

We recognize that our work has several limitations. While thorough, decades of work in disagreement make a fully exhaustive survey difficult. Our survey reviews over 120 papers, and our selection reflects methodological relevance to disagreement-aware modeling. As a result, some related approaches, such as LLM simulations of annotator personas, or other methods that engage with human perspectives tangentially rather than as a modeling target, have been excluded. Additionally, the taxonomy we propose is a simplification. In real-world settings, data, task, and annotator factors sometimes cannot be separated cleanly. Future approaches, particularly those that integrate the interaction between various factors, might require extensions to this taxonomy. Lastly, our discussion of open challenges and future directions reflects our interpretations of the most important gaps in the literature. These should not be interpreted as a comprehensive or normative research agenda. Other researchers might prioritize different directions, such as efficiency, model deployment concerns, or annotation practice design to produce meaning disagreement.

A key limitation of the disagreement-aware literature (and thus our synthesis) is that disagreement is often treated as an observed property of datasets rather than as an outcome of annotation practice design. Observed disagreement may conflate meaningful perspective variation with annotator factors such as confusion and fatigue. This limits cross-paper comparability. Developing and standardizing annotation practices that intentionally elicit meaningful disagreement remains an important direction for future work.

References

Gavin Abercrombie, Tanvi Dinkar, Amanda Cercas Curry, Verena Rieser, and Dirk Hovy. 2025. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *Preprint*, arXiv:2301.10684.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta. Association for Computational Linguistics.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Reuben Binns. 2018. Fairness in algorithmic decision-making: Lessons from political philosophy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 149–159. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Z. Cao, E. Chen, Y. Huang, S. Shen, and Z. Huang. 2023. Learning from crowds with annotation reliability. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2103–2107.

Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

Mauro Cazzaniga, Florence Jaumotte, Longji Li, Giovanni Melina, Augustus J. Panton, Carlo Pizzinelli, Emma J. Rockall, and Marina Mendes Tavares. 2024. Gen-ai: Artificial intelligence and the future of work. Technical Report 2024/001, International Monetary Fund.

Jonathan P. Chang, Saleema Amershi, and Jaime Teevan. 2017. Revolt: Collaborative crowdsourcing for labeling tasks. In *Proceedings of CHI*.

Z. Chu, J. Ma, and H. Wang. 2021. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5832–5840.

Berta Chulvi, Lara Fontanella, Roberto Labadie Tamayo, and Paolo Rosso. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *NLPerspectives@ECAI*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2007–2021, New York, NY, USA. Association for Computing Machinery.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*, pages 469–478, Lyon, France. Association for Computing Machinery.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Russel Dsouza and Venelin Kovatchev. 2025. Sources of disagreement in data for LLM instruction tuning. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 20–32, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *Preprint*, arXiv:1808.06080.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *CoRR*, abs/2306.16388.

Jing Fan, Guoliang Li, Beng Chin Ooi, Kian-Lee Tan, and Jun Feng. 2015. icrowd: An adaptive crowdsourcing framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1015–1030, Melbourne, Victoria, Australia. Association for Computing Machinery.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Simona Frenda, Valerio Basile, Tommaso Caselli, and Viviana Patti. 2024. Perspectivist approaches to natural language processing: A survey. *Natural Language Processing and Artificial Intelligence*. Survey paper.

Mirta Galesic and Michael Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2):349–360.

Z. Gao, F.-K. Sun, M. Yang, S. Ren, Z. Xiong, M. Engeler, A. Burazer, L. Wildling, L. Daniel, and D. S. Boning. 2022. Learning from multiple annotator noisy labels via sample-wise label fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 407–422. Springer.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller. 2023. Crowdlab: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. *Preprint*, arXiv:2210.06812.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, page 1–19. ACM.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

M. Guan, V. Gulshan, A. Dai, and G. Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tao Han, Huaixuan Shi, Xinyi Ding, Xiao Ma, Huamao Gu, and Yili Fang. 2025. Mixture of experts based multi-task supervise learning from crowds. *Preprint*, arXiv:2407.13268.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

Daniil Ignatev, Denis Paperno, and Massimo Poesio. 2025. Hypernetworks for perspectivist adaptation. *Preprint*, arXiv:2510.13259.

Oana Inel, K Khamkham, T Cristea, A Dumitrache, H Rutjes, R-J Sips, J van Ossenbruggen, and K Crowston. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web – ISWC 2014*, volume 8797 of *Lecture Notes in Computer Science*, pages 486–504, Cham. Springer.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. Nutmeg: Separating signal from noise in annotator disagreement. *Preprint*, arXiv:2507.18890.

Uthman Jinadu and Yi Ding. 2024. Noise correction on subjective datasets. *Preprint*, arXiv:2311.00619.

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

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Jon A. Krosnick, Sowmya Narayan, and Wendy R. Smith. 1996. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44.

Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. 2015. A bayesian framework for modeling human evaluations. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*, pages 181–189.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Qiang Liu, Jian Peng, and Alexander T. Ihler. 2012. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems*, volume 25, pages 692–700.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. PERSEVAL: A framework for perspectivist classification evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22345–22370, Suzhou, China. Association for Computational Linguistics.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2021. Detecting stance in media on global warming. *Preprint*, arXiv:2010.15149.

Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings*

of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 745–754, Sydney, NSW, Australia. Association for Computing Machinery.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Woźniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocoń. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *NLPerspectives@ECAI*.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Christoph Molnar. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com.

Britta Mondorf. 2002. Gender differences in english syntax. *Journal of English Linguistics*, 30(2):158–180.

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

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *Preprint*, arXiv:2502.20897.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Jahna Otterbacher. 2018. Social cues, social biases: Stereotypes in annotations on people images. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 136–144.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20333–20349, Suzhou, China. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS-52)*.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018a. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018b. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018c. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inference. In *Proceedings of ACL*.

Barbara Plank, Hans Van Halteren, and Anders Søgaard. 2014. Learning part-of-speech tagging with partial annotations. In *Proceedings of COLING*, pages 483–492.

Massimo Poesio. 2020. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Companion to Semantics*. Wiley.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

W. S. Robinson. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357.

F. Rodrigues and F. Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1611–1618.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Yisi Sang and Jeffrey Stanton. 2021. The origin and value of disagreement among data labelers: A case study of the individual difference in hate speech annotation. *Preprint*, arXiv:2112.04030.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 29971–30004, Honolulu, Hawaii, USA. PMLR.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Edgar D. Simpson, Matteo Venanzi, Pushmeet Kohli, John Guiver, Gianluca Kazai, and Milad Shokouhi. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*, pages 992–1002, Florence, Italy. Association for Computing Machinery.

Fritz Strack. 1992. "order effects" in survey research: Activation and information functions of preceding questions. In Norbert Schwarz and Seymour Sudman, editors, *Context Effects in Social and Psychological Research*, chapter 3. Springer, New York, NY.

R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman. 2019. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11244–11253.

Alexandra Uma, Tommaso Fornaciari, Silviu Paun, Benjamin Paul Chamberlain, Dirk Hovy, Barbara Plank, and Dean K Simonton. 2021. Learning from disagreement: A survey. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 9, pages 1408–1424.

Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 111–115, Torino, Italia. ELRA and ICCL.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

H. Wei, R. Xie, L. Feng, B. Han, and B. An. 2022. Deep learning from multiple noisy annotators as a union. *IEEE Transactions on Neural Networks and Learning Systems*.

Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, volume 23, pages 2424–2432.

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

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Conference on Natural Language Processing*.

Jin Xu, Mariët Theune, and Daniel Braun. 2024a. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Shanshan Xu, T. Y. S. S Santosh, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2024b. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. *Preprint*, arXiv:2310.11878.

Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. 2025. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. *Preprint*, arXiv:2508.02853.

Tianling Yang, Christian Strippel, Alexandra Keiner, Dylan Baker, Alexis Chávez, Krystal Kauffman, Marc Pohl, Caroline Sinders, and Milagros Miceli.

2023. Ethics of data work: Principles for academic data work requesters. Weizenbaum discussion paper, Weizenbaum Institute for the Networked Society.

Liyun Zhang, Jingcheng Ke, Shenli Fan, Xuanmeng Sha, and Zheng Lian. 2025a. A unified evaluation framework for multi-annotator tendency learning. *Preprint*, arXiv:2508.10393.

Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2025b. Diverging preferences: When do annotators disagree and do models know?

Zhao Zhao, Fang Wei, Ming Zhou, Wei Chen, and Wilfred Ng. 2015. Crowd-selection query processing in crowdsourcing databases: A task-driven approach. In *Proceedings of the 18th International Conference on Extending Database Technology (EDBT)*, pages 397–408. OpenProceedings.org.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *Proc. VLDB Endow.*, 10(5):541–552.

A Appendix

A.1 Survey Scope and Selection

This survey focuses on research in NLP that studies, models, or evaluates annotator disagreement in text-based tasks. We restrict our scope to NLP settings where disagreement is theoretically meaningful, such as toxicity detection, stance analysis, sentiment, and other subjective or socially grounded judgments, rather than domains where disagreement primarily reflects measurement error. Our goal is not to exhaustively catalog all work on annotation variability (of which there is decades of work), but to synthesize lines of research that treat disagreement as a signal relevant to learning, evaluation, or model design, with a slight focus on connecting the most recent work.

We survey over 120 prior works spanning disagreement sources, modeling methods, and evaluation paradigms. Table 1 provides the counts of papers surveyed in each category we focus on. Papers may appear in multiple categories when relevant. We also reviewed approximately 12 papers on large language model–based persona prompting and demographic conditioning; however, because most of this work focuses on simulating personas rather than explicitly modeling annotator structure or disagreement, we discuss these primarily in the Challenges and Future Work section.

We include work that makes a substantive methodological contribution, defined as proposing a learning algorithm, prediction target, or evaluation framework that explicitly accounts for multiple annotators or divergent interpretations. For each of the three modeling families that we identify (latent truth modeling, task-based multi-annotator learning, and embedding-based approaches), we emphasize seminal and highly-cited work (Dawid and Skene, 1979; Cohn and Specia, 2013; Kocoń et al., 2021), alongside representative recent advances. Our taxonomy of disagreement sources was developed iteratively in tandem with the literature review, with theoretical distinctions refined through engagement with empirical findings.

Survey Category	**# Papers**
Sources of Disagreement	25
Latent Truth & Reliability Modeling	22
Task-Based Multi-Annotator Learning	12
Embedding-Based Multi-Annotator Learning	15
Evaluation Metrics	18
Total Unique Papers Surveyed	120

Table 1: Distribution of papers surveyed across categories. Papers may appear in multiple categories.

Taken together, this scope allows us to position our survey as a complement to existing work. Whereas prior surveys are either primarily conceptual (Frenda et al., 2024) or modality-general (Uma et al., 2021), we center NLP as a domain where disagreement is deeply tied to task design, interpretation practices, and annotator identities. By jointly examining modeling approaches, evaluation practices, and their fairness implications within a unified framework, we aim to provide a more technically-grounded synthesis of perspectivist modeling that has not yet been articulated in existing surveys. Finally, to capture emerging trends, we closely follow work from the Workshop on Perspectivist Approaches to NLP[1] and related venues.

A.2 Detailed Synthesis Table of Prediction Targets and Pooling Structures

We provide a detailed table mapping representative work in each modeling family to the axis of prediction target and pooling structures (Fig. 4).

[1] https://nlperspectives.di.unito.it/

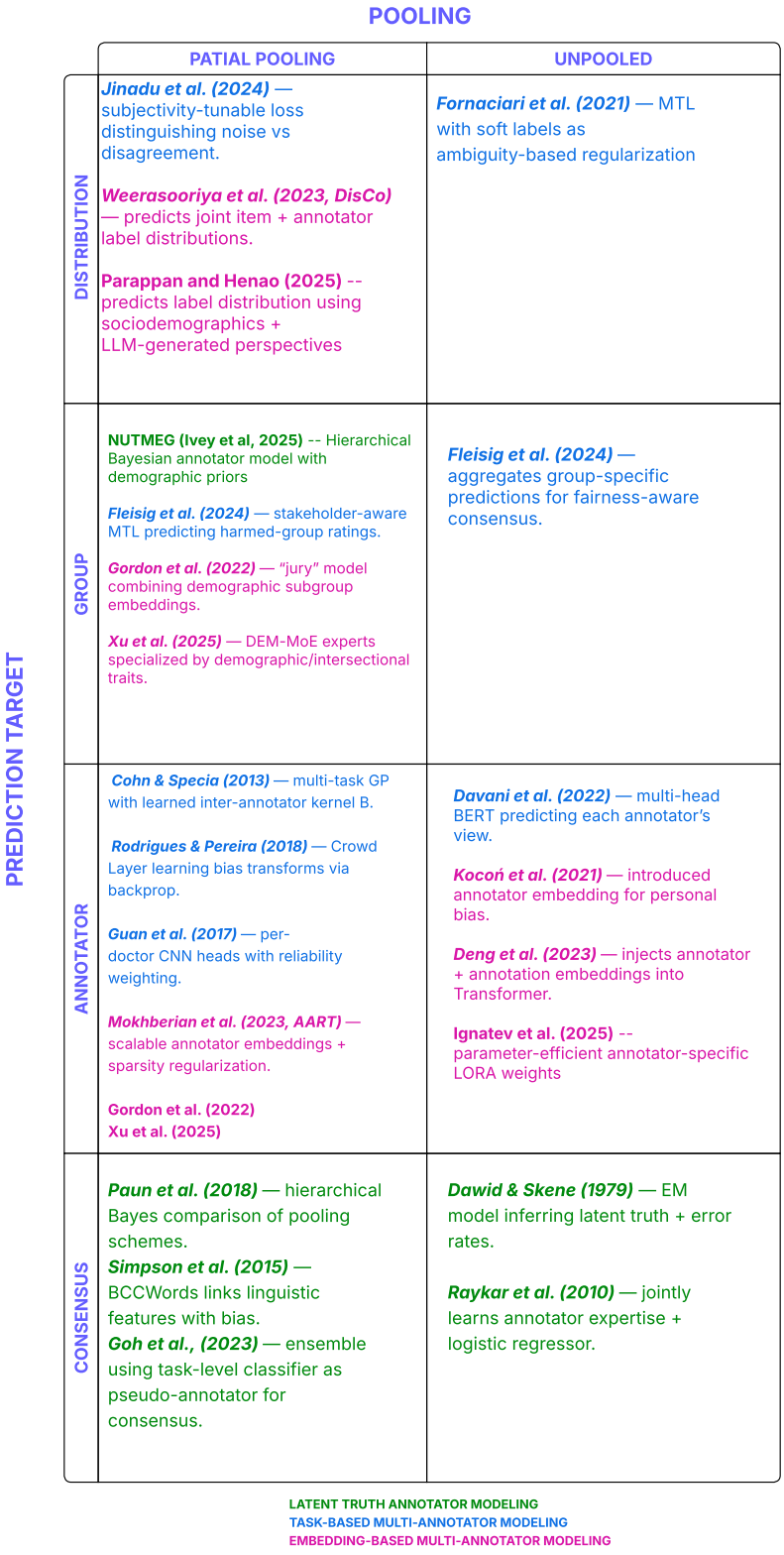

Figure 4: Prediction targets and pooling assumptions across methods for learning from annotator disagreement. The table maps existing approaches by what they predict (consensus labels, individual annotator responses, group-level outputs, or full disagreement distributions) and how they pool information across annotators. Modeling choices implicitly encode different assumptions about population structure, perspective aggregation with recent work increasingly favoring partial pooling and distributional targets.