Bias in, Bias out: Annotation Bias in Multilingual Large Language Models

Xia Cui[1], Ziyi Huang[2], Naeemeh Adel[1]
[1]Manchester Metropolitan University, Manchester, UK.
{x.cui, n.adel}@mmu.ac.uk
[2]Hubei University, Wuhan, China.
ziyihuang@hubu.edu.cn

Abstract

Annotation bias in NLP datasets remains a major challenge for developing multilingual Large Language Models (LLMs), particularly in culturally diverse settings. Bias from task framing, annotator subjectivity, and cultural mismatches can distort model outputs and exacerbate social harms. We propose a comprehensive framework for understanding annotation bias, distinguishing among *instruction bias*, *annotator bias*, and *contextual and cultural bias*. We review detection methods (including inter-annotator agreement, model disagreement, and metadata analysis) and highlight emerging techniques such as multilingual model divergence and cultural inference. We further outline proactive and reactive mitigation strategies, including diverse annotator recruitment, iterative guideline refinement, and post-hoc model adjustments. Our contributions include: (1) a structured typology of annotation bias, (2) a comparative synthesis of detection metrics, (3) an ensemble-based bias mitigation approach adapted for multilingual settings, and (4) an ethical analysis of annotation processes. Together, these contributions aim to inform the design of more equitable annotation pipelines for LLMs.

1 Introduction

Large Language Models (LLMs) such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), Llama (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023) have transformed Natural Language Processing (NLP), achieving state-of-the-art performance across a wide range of tasks. Their success is largely attributed to pre-training on vast, unannotated corpora that enable them to learn powerful representations. However, aligning these models with human values and adapting them for high-stakes applications requires smaller, curated datasets annotated by humans.

This reliance introduces a critical vulnerability. Annotation bias, which refers to systematic distortions introduced during the labelling process, can severely affect model performance, fairness, and generalisation. It may arise from task framing, annotator subjectivity, or cultural mismatches, and its impact is particularly pronounced in multilingual and culturally heterogeneous contexts (Bender and Friedman, 2018; Plank, 2022).

The consequences of annotation bias are not hypothetical. For example, models trained to detect toxicity often misclassify African-American Vernacular English (AAVE) as offensive, due to cultural insensitivity in both annotation guidelines and annotator interpretation. Phrases such as "That's my nigga" which carry a supportive meaning in AAVE, are frequently labelled as hateful by annotators unfamiliar with the dialect (Sap et al., 2019). This highlights how linguistic and cultural assumptions embedded in the annotation process can lead to unjust model behaviour.

Such failures reflect a broader pattern. When biased annotations are used for training or fine-tuning, models tend to replicate and sometimes amplify these distortions, resulting in both representational harms and disparities in performance across demographic groups (Dodge et al., 2021; Sheng et al., 2019). Addressing these issues requires critical scrutiny of annotation workflows, with careful attention to cultural and contextual diversity.

In this paper, we examine the sources and consequences of annotation bias in multilingual LLMs. We propose a typology of annotation bias, encompassing instruction bias, annotator bias, and contextual or cultural bias. We review established and emerging detection methods, including inter-annotator agreement, model disagreement, and multilingual divergence. We adapt Weak Ensemble Learning (WEL) as a reactive mitigation strategy

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 1–16, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-001

and assess its effectiveness across multilingual and real-world datasets. Finally, we reflect on the ethical and labour implications of annotation work and suggest directions for building more inclusive and context-aware NLP pipelines.

2 Background and Motivation

Early annotation practices in NLP were shaped by linguistic theory and typically involved trained experts using detailed, rule-based guidelines. Datasets such as the Penn Treebank (Marcus et al., 1993) and FrameNet (Baker et al., 1998) exemplified this approach, producing consistent annotations at a small to moderate scale.

As NLP tasks expanded and model complexity increased, the field shifted toward large-scale annotation through crowdsourcing platforms (Snow et al., 2008a). This approach enabled the creation of widely used datasets like SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), but introduced new concerns regarding annotation quality, consistency and subjectivity.

The rise of LLMs has further complicated annotation workflows. Today's datasets often combine expert review, crowdworker input, and semi-automated methods such as model-in-the-loop annotation (Pawar et al., 2025). Many target inherently subjective or ambiguous constructs, including helpfulness, safety, or moral alignment (Monarch, 2021; Uma et al., 2022). These tasks are particularly vulnerable to variation across annotators and contexts.

At the same time, NLP has increasingly embraced multilingual and multimodal benchmarks. Projects such as XTREME (Hu et al., 2020), AmericasNLI (Ebrahimi et al., 2022), TextVQA (Singh et al., 2019), and HowTo100M (Miech et al., 2019) highlight the challenges of applying traditional annotation schemas across languages, cultures, and modalities. Multimodal tasks introduce further complexity through temporality, affective signals, and cross-modal interpretation.

These trends have exposed a structural vulnerability: annotation bias. This includes not only individual annotator subjectivity, but also culturally conditioned assumptions, linguistic mismatches, and platform-mediated incentives (Blodgett et al., 2020; Plank, 2022). Annotation decisions made on small but influential datasets can propagate through model fine-tuning and evaluation, leading to downstream harms (Dodge et al., 2021).

In response, the field has developed ethical documentation frameworks such as *Data Statements* (Bender and Friedman, 2018) and *Datasheets for Datasets* (Gebru et al., 2021). These initiatives promote transparency by capturing the dataset's linguistic, demographic, and procedural context. They represent an important step toward recognising that high-quality, ethical NLP systems begin with well-understood and well-documented data.

3 Types of Annotation Bias

Annotation bias in NLP arises when human interpretations, cultural assumptions, or task formulations systematically distort labelled data. These biases can affect model learning, especially during fine-tuning and evaluation. In the context of LLMs, annotation bias often originates from multiple sources and compounds over the pipeline. Addressing it requires distinguishing among different types of bias and understanding how they interact.

We categorise annotation bias into three primary types based on its origin: **Instruction Bias** (Section 3.1), **Annotator Bias** (Section 3.2), and **Contextual and Cultural Bias** (Section 3.3). These types are not mutually exclusive. In many cases, a biased annotation reflects an interaction among all three. For example, culturally narrow task guidelines (instruction bias) given to a homogeneous annotator pool (annotator bias) tasked with labelling dialectal language (contextual bias) may produce systematically skewed data. Recognising this interplay is essential for designing effective detection and mitigation strategies (Bender and Friedman, 2018).

3.1 Instruction Bias

Instruction bias (Parmar et al., 2023) occurs when the design of an annotation task, including the prompt wording, labelling guidelines, or interface, embeds implicit assumptions that shape how annotators interpret or respond. These assumptions can systematically distort the resulting labels.

A common example appears in sentiment analysis, where annotators are often asked to classify texts as "positive", "negative", or "neutral". These categories overlook cultural and linguistic nuance, such as expressions of irony, ambivalence, or indirectness (Mohammad, 2016). The framing tends to reflect Western emotional norms that do not generalise across diverse populations (Huang and Yang,

2023). Similarly, toxicity detection tasks have been shown to mislabel minoritised dialects as offensive, in part due to annotation instructions that lack sociolinguistic sensitivity (Sap et al., 2019).

In the context of LLMs, instruction bias is further complicated by the use of prompts in place of formal annotation guidelines. *Zero-shot* (Wei et al., 2022) and *few-shot* prompting (Schick and Schütze, 2022) methods often replace expert-designed protocols. These prompts, though brief, function as implicit task instructions and strongly influence model behaviour. Minor changes in phrasing, such as asking "Is this inappropriate?" versus "Is this morally wrong?", can lead to significantly different model outputs, especially for subjective or value-laden tasks (Zhao et al., 2021; Schick and Schütze, 2022; He et al., 2024).

Moreover, prompts are frequently written by researchers or practitioners who come from specific cultural or disciplinary contexts. Their assumptions shape how tasks are framed and what kinds of answers are considered valid. For example, in mental health detection tasks, prompt templates often reflect Western norms of psychological distress. This reduces model performance on data from underrepresented linguistic or cultural groups (Parmar et al., 2023; Cui et al., 2024). Unlike traditional annotation guidelines, prompts are rarely revised or reviewed through participatory validation processes (Zamfirescu-Pereira et al., 2023; Cui et al., 2024).

3.2 Annotator Bias

Annotator bias arises from the individual or collective predispositions of those performing the labelling. These may include cognitive heuristics, beliefs, social norms, or demographic characteristics. Even when given identical instructions, annotators interpret data differently depending on their personal context.

Subjective annotation tasks such as toxicity detection, moral judgment, or hate speech classification are particularly susceptible to this type of bias (Sap et al., 2019; Liu et al., 2022; Plank, 2022). Aggregation techniques like majority voting can obscure these differences and suppress minority perspectives, especially when annotator diversity is limited (Aroyo and Welty, 2015; Shardlow, 2022).

The rise of crowdwork has intensified these challenges. Annotator pools often differ demographically from both the dataset's source community and its intended application domain (Eickhoff, 2018; Bender et al., 2021). As a result, annotations may misinterpret cultural cues, dialectal language, or context-specific emotional tone. Although such variation is not necessarily the result of carelessness, it can introduce systematic distortion, especially when disagreement is treated as noise rather than signal (Cabitza et al., 2023).

3.3 Contextual and Cultural Bias

Contextual and cultural bias occurs when task design and labelling decisions assume a particular worldview, linguistic norm, or social context. It becomes especially pronounced in multilingual and multimodal tasks, where language, meaning, and affective signals vary widely across cultures.

Annotation labels such as "polite", "supportive", or "offensive" often fail to translate cleanly across languages or communities (Ponti et al., 2020). Cultural norms shape how people interpret both language and non-verbal cues, including gestures and tone of voice (Barrett et al., 2019; Lukac et al., 2023). Without regionally grounded interpretation frameworks, annotators may mislabel visual or emotional content.

Additionally, most pretraining data is skewed toward English and Western sources. As of 2025, English accounts for nearly half of all indexed web content (Ani Petrosyan, 2025). This imbalance in data collection reinforces a corresponding bias in annotation practices.

Recent work has emphasised the importance of culturally grounded taxonomies and community consultation for annotation tasks involving identity, emotion, or morality (Blodgett et al., 2020; Hutchinson et al., 2020; Zhou et al., 2023). Without such grounding, models trained on annotated data risk reproducing narrow, non-representative worldviews.

4 Impact on Model Behaviour

Bias introduced during annotation does not remain confined to the dataset. It propagates into the models trained on that data and leads to measurable downstream harms. This phenomenon, often referred to as "bias in, bias out," is a central concern in machine learning. When annotation processes reflect cultural, social, or demographic distortions, models tend to reproduce those distortions, and in some cases, amplify them (Dodge et al., 2021).

One of the most well-documented consequences

is performance disparity across demographic groups. A model may perform well on aggregate metrics while underperforming on texts associated with certain identities, dialects, or cultural contexts. For example, commercial gender classification systems have shown much higher error rates for darker-skinned women. This discrepancy can be traced, in part, to unbalanced training data that lacked diverse and properly annotated examples (Buolamwini and Gebru, 2018). Similarly, recidivism prediction tools have displayed racially skewed false positive rates due to historical biases embedded in the labelled data (Dressel and Farid, 2018).

Beyond accuracy gaps, annotation bias also causes representational harm. These occur when models learn to reproduce social stereotypes or unfair associations. For instance, if training labels disproportionately associate "engineer" with men and "nurse" with women, the model may internalise and repeat these biases in downstream tasks such as text generation or summarisation (Sheng et al., 2019). In a similar way, toxicity detection models trained on biased annotations may misclassify expressions in African-American Vernacular English (AAVE) as hostile or inappropriate (Sap et al., 2019).

These harms can be formalised using established fairness metrics. **Demographic Parity** requires that the rate of positive predictions be equal across groups. **Equalised Odds** requires that true and false positive rates remain consistent regardless of group membership. Annotation bias undermines these goals. Returning to the AAVE example, if annotators are more likely to label AAVE expressions as toxic, a classifier trained on such data will exhibit a higher false positive rate for Black speakers. This violates Equalised Odds and leads to unfair penalties against specific communities (Dixon et al., 2018).

These examples demonstrate that annotation bias is not a peripheral issue. It directly contributes to systemic failures in LLMs that affect both technical performance and social impact. For this reason, examining and improving annotation practices is a foundational step toward fairer and more reliable NLP systems.

5 Case Studies and Empirical Evidence

To illustrate how annotation bias operates in practice, this section presents two case studies. The first addresses multilingual hate speech detection, where cultural definitions of offence lead to misalignment between training data and deployment contexts. The second focuses on multimodal emotion recognition, where non-verbal cues are interpreted differently across cultural frameworks. These cases highlight that bias often arises not from individual prejudice but from structural mismatches between annotation design and communicative diversity.

5.1 Case Study: Cross-Cultural Hate Speech Detection

Hate speech detection is highly sensitive to cultural context. What is considered offensive or harmful in one setting may be acceptable or even humorous in another. This presents a serious challenge for creating models intended to generalise across regions and languages.

Lee et al. (2023) evaluated monolingual hate speech classifiers across cultural contexts by applying models trained on English-language data from the United States to translated data from languages such as Korean and Arabic. The results showed a drop in F1 scores of up to 42% and a fourfold increase in false negatives. These failures stemmed not from technical flaws in the models themselves but from annotation biases embedded in the source datasets. Several factors contributed to this performance collapse:

- Cultural targets vary. Groups and individuals who are frequent targets of hate speech differ between cultures, meaning training data from one country may miss important examples from another.
- Sociocultural norms shape expression. Sarcasm, irony, and rhetorical devices have different meanings and social functions depending on the culture.
- Standards of offensiveness diverge. A statement considered hateful in one community may be seen as neutral or even acceptable in another, depending on social, political, or historical context.

This case demonstrates that hate speech is not a culturally neutral construct. Models built on datasets annotated within a single cultural context may fail when applied elsewhere, even if the language is translated. This failure is not only a limitation of model generalisation but also a direct consequence of annotation bias in the original data.

5.2 Case Study: Multimodal Emotion Recognition

Bias in multimodal datasets can be more difficult to detect but equally damaging. Emotion recognition tasks that use audio, video, or gesture data rely on the interpretation of non-verbal cues, which are deeply culturally embedded.

Gunes and Piccardi (2007) conducted a study where they investigated how physical gestures were interpreted across cultural contexts. They found that a single gesture could signal patience in Egypt, positivity in Greece, and confrontation in Italy. When such data are annotated by individuals unfamiliar with the cultural origin of the gesture, systematic mislabelling is likely.

Cultural variation also affects emoji and facial expression interpretation. Gao and VanderLaan (2020) showed that annotators from Western cultures rely more on mouth shapes to read emoji emotions, while those from Eastern cultures prioritise the eyes. These perceptual differences result in inconsistent annotations and affect model training when emojis are used as supervision signals.

These findings underscore the importance of culturally grounded annotation frameworks in multimodal NLP. Without them, datasets risk encoding a narrow view of human emotion and interaction, reducing the validity and generalisability of trained models.

6 Detecting Annotation Bias

Detecting annotation bias is a crucial step toward mitigating its impact on model training and evaluation. A variety of methods have been proposed to identify systematic patterns of bias in annotated datasets, each with different strengths and limitations. One common approach is to measure inter-annotator agreement (IAA), using metrics such as Cohen's κ (Smeeton, 1985). For N instances and M annotators, where $y_i^{(j)} \in \mathcal{Y}$ is annotator j's label for instance i, the agreement is defined as:

$$\kappa_{\text{Cohen}} = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i^{(1)} = y_i^{(2)}) \tag{1}$$

Here p_o is the observed agreement, i.e., the probability that both annotators assign the *same* label to a randomly selected item. $\mathbb{I}(\cdot)$ is the indicator function. The expected agreement by chance p_e is computed as:

$$p_e = \sum_{k \in \mathcal{Y}} P(y^{(1)} = k) \cdot P(y^{(2)} = k) \tag{2}$$

where $P(y^{(j)} = k)$ is the empirical probability of annotator j assigning label k. Fleiss' κ generalises this metric for multiple annotators (Fleiss et al., 2013):

$$\kappa_{\text{Fleiss}} = \frac{\bar{p} - \bar{p}_e}{1 - \bar{p}_e}, \bar{p} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{k=1}^{K} n_{ik}(n_{ik} - 1)}{M(M-1)} \tag{3}$$

where n_{ik} is the number of annotators assigning label k to instance i.

For settings with missing data or mixed label types, Krippendorff's α (Krippendorff, 2011) offers a more general reliability metric:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{4}$$

where D_o is the observed disagreement (weighted across annotator pairs per item) and D_e is the expected disagreement under chance.

A complementary approach is to analyse *model disagreement*. When two models are trained on the same data, divergence in their predictions can reveal annotation ambiguity or bias (Geva et al., 2019). For two models f_1 and f_2, the disagreement rate (DR) is defined as:

$$\text{DR} = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} \mathbb{I}\big(f_1(x) \neq f_2(x)\big) \tag{5}$$

Uma et al. (2022) extend this idea by comparing model predictions with human labels:

$$\Delta = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} \big|f(x) - y_{\text{human}}(x)\big| \tag{6}$$

This metric helps identify inconsistencies between model behaviour and annotation patterns (Dsouza and Kovatchev, 2025).

Another lens on bias detection comes from **metadata analysis**. By examining annotator demographics, task context, and label distributions, researchers can uncover systematic bias (Sap et al., 2019). For an annotator group a, a demographic gap $G(a)$ can be computed as:

$$G(a) = \left| \frac{1}{|\mathcal{D}_a|}\sum_{x \in \mathcal{D}_a} y(x) - \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} y(x) \right| \tag{7}$$

Here, $\mathcal{D}_a$ denotes the subset of data annotated by group a, and $y(x)$ is the label assigned to instance x. A high $G(a)$ may signal systematic differences in annotation patterns between group a and the overall dataset, potentially reflecting underlying biases or cultural variation.

This gap measures how far the group's average labels deviate from the global average, which may indicate bias or representational disparity (Sap et al., 2019). Traditional metrics, however, may be less effective in multilingual and culturally diverse settings. In these cases, disagreement may reflect true variation rather than annotation error (Naous et al., 2024). To address this, new strategies are emerging. *multilingual model disagreement* compares the predictions of models fine-tuned in different languages on parallel corpora $\mathcal{D}_{l_1,l_2}$:

$$\mathrm{DR}(l_1, l_2) = \frac{1}{|\mathcal{D}_{l_1,l_2}|} \sum_{x \in \mathcal{D}_{l_1,l_2}} \mathbb{I}\big(f_{l_1}(x) \neq f_{l_2}(x)\big) \tag{8}$$

where f_{l_1} and f_{l_2} denote the models fine-tuned in languages l_1 and l_2, respectively.

Similarly, cultural inference techniques (Zhang et al., 2020b; Huang and Yang, 2023) use embeddings or sociolinguistic metadata to detect alignment between annotations and cultural backgrounds. One such indicator, Φ_{cultural} is calculated as the ℓ_2 distance between two groups:

$$\Phi_{\text{cultural}} = \| \phi(\mathcal{D}_a) - \phi(\mathcal{D}_{a'}) \|_2 \tag{9}$$

where $\phi(\cdot)$ maps a dataset to its cultural embedding space, and $\mathcal{D}_a$, $\mathcal{D}_{a'}$ denote datasets annotated by cultural groups a and a', respectively.

Together, these methods offer a toolkit for identifying annotation bias at different levels: label consistency, annotator disagreement, cultural framing, and model interpretation. In practice, combining quantitative metrics with qualitative analysis offers the best chance of uncovering and addressing complex forms of annotation bias.

7 Mitigation Strategies

Detecting annotation bias is only the first step toward creating fair and reliable NLP systems. Effective mitigation requires both proactive strategies, which aim to prevent bias during data collection, and reactive strategies, which address it after annotation or model training. This section outlines techniques across both categories, integrating recent formal approaches with practical best practices.

7.1 Proactive Strategies

Proactive strategies aim to reduce annotation bias at the source by redesigning annotation processes with awareness of potential pitfalls.

Diverse Annotator Pools To counter annotator bias, it is essential to recruit annotators from a broad range of demographic, cultural, and linguistic backgrounds (Bender et al., 2021; Paullada et al., 2021). A diverse pool can reveal meaningful disagreements and represent underreported perspectives (Aroyo and Welty, 2015). One way to quantify diversity is through the entropy of the demographic distribution:

$$H(A) = -\sum_{a \in \mathcal{A}} p(a) \log p(a) \tag{10}$$

where $\mathcal{A}$ is the set of annotator groups and $p(a)$ is the proportion of annotations from group a. A higher entropy score $H(A)$ indicates a more balanced and inclusive annotation pool.

Dynamic Annotation Guidelines To mitigate instruction bias, guidelines and prompts should be piloted, reviewed and refined iteratively. This feedback loop helps remove culturally specific assumptions and linguistic ambiguities (Parmar et al., 2023). In LLM-based settings, prompt engineering should be evaluated across cultural contexts to ensure validity (Zamfirescu-Pereira et al., 2023). One can formalise this iterative process by tracking the variance in annotator disagreement across iterations:

$$\sigma_t^2 = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathrm{Var}\big(\{y_i^{(t)}(x)\}_{i=1}^n\big) \tag{11}$$

where $y_i^{(t)}(x)$ is the label from annotator i on item x during iteration t, with the goal that $\sigma_t^2 \to \min$ over t.

Culturally Grounded Taxonomies To address contextual and cultural bias, annotation schemes should be developed with culturally grounded taxonomies of emotion, politeness, morality, and related constructs (Blodgett et al., 2020; Hutchinson et al., 2020; Zhou et al., 2023). Engaging with communities or domain experts helps ensure that annotation labels are valid across languages and cultural settings (Ponti et al., 2020; Naous et al., 2024).

7.2 Reactive Strategies

Reactive strategies are applied after biases have entered the data or the model. They aim to mitigate downstream harms without necessarily revising the annotation process itself. One key challenge in post-hoc mitigation is handling inconsistencies introduced by annotator subjectivity and instruction bias, particularly when labels reflect divergent interpretations of subjective or culturally loaded concepts. Weighted ensemble methods can address this by leveraging multiple model perspectives to smooth over annotation noise, while still preserving minority viewpoints.

Post-hoc Model Adjustment Biases in trained models can sometimes be mitigated using post-hoc correction methods such as embedding debiasing or output regularisation. Kaneko et al. (2023) proposed modifying model outputs by subtracting a learned bias component:

$$f^{\text{debias}}(x) = f_\theta(x) - \lambda b(x) \tag{12}$$

where $f_\theta(x)$ is the original model output, $b(x)$ is a bias projection and λ controls debiasing strength.

Fine-tuning and In-context Debiasing Recent work has explored using targeted fine-tuning (Webster et al., 2021) or in-context prompting (Ganguli et al., 2023) to reshape model behaviour without altering the training data (Kaneko et al., 2025). In the fine-tuning case, model parameters θ are updated using reweighted or re-annotated dataset $\mathcal{D}^*$:

$$\min_\theta \mathbb{E}_{(x,y^*)\sim\mathcal{D}^*}\mathcal{L}(f_\theta(x), y^*) \tag{13}$$

In in-context learning, models are conditioned on carefully constructed prompts P that reduce bias:

$$f_\theta(y \mid x, P) \tag{14}$$

where P is designed to reduce the likelihood of biased completions while preserving task accuracy.

Multi-Objective Weighted Ensemble Learning Another reactive strategy leverages ensemble learning to mitigate annotation bias by explicitly modelling annotator disagreement (Geva et al., 2019). Given a dataset $\mathcal{D} = \{(x_i, \{y_i^{(j)}\}_{j=1}^M)\}_{i=1}^N$, Huang et al. (2025) proposed Weak Ensemble Learning (WEL), which samples one annotator label per instance to construct K label-variant datasets. Each trains a weak predictor f_{θ_k}, weighted by its held-out performance (e.g., F1, cross-entropy, Manhattan distance), with $\sum_{k=1}^K w_k = 1$. We extend WEL to a multilingual setting by applying the same label-sampling procedure across datasets in different languages using a shared multilingual model. Final predictions are computed as:

$$\hat{y}_i = \sum_{k=1}^{K} w_k \, f_{\theta_k}(x_i), \tag{15}$$

allowing the ensemble to capture annotator disagreement while leveraging multilingual representations from a single model.

We use `mBERT` (Devlin et al., 2019) as the base model. On the multi-source benchmark from the LeWiDi 2023 shared task (Leonardellli et al., 2023), WEL generally outperforms baselines using single-model CE loss (CE-only) (Uma et al., 2020) and majority-vote ensembles of top five annotators (Top-5-Ann) (Xu et al., 2024), achieving higher F1 and lower CE/MD scores. The only exception is *ArMIS*, where the very small annotator pool (three annotators) limits the effectiveness of random label sampling. As the primary focus of this paper is on the discussion of annotation bias in multilingual LLMs, we include the full experimental results in Appendix B.

8 Ethical and Practical Considerations

The discussion of annotation bias is incomplete without considering the ethical and practical realities of the annotation process itself. Creating high-quality labelled data is not only a technical challenge but also a form of labour that carries human and institutional consequences. These concerns are directly tied to the emergence and persistence of annotation bias because they influence how data are produced, who produces it, and under what conditions.

8.1 Annotator Wellbeing and Psychological Safety

One of the most pressing concerns involves the well-being of annotators, particularly those responsible for labelling harmful, toxic, or distressing content. Content moderation datasets, which are essential for training safety filters in LLMs, often expose annotators to a continuous stream of violent, hateful, or traumatic material. Research shows that prolonged exposure to such content can lead to severe psychological effects, including anxiety, depression, insomnia, and symptoms of post-traumatic stress disorder (PTSD) (Das et al., 2020).

This phenomenon is referred to as *vicarious trauma*, a condition in which individuals who are indirectly exposed to trauma begin to show symptoms similar to those of direct trauma survivors (Pearlman and Saakvitne, 1995). These effects are compounded by stressful work environments. Annotators often face tight deadlines and high task volumes, with limited autonomy or support systems (Spence et al., 2023). In many cases, stigma around mental health further prevents them from seeking help (Bergman and Rushton, 2023).

To mitigate these harms, researchers and data curators have a responsibility to implement safeguards. These may include access to mental health services, task rotation to reduce exposure to distressing material, and policies that allow annotators to opt out of specific assignments. Regular breaks, content warnings, and workplace cultures that promote psychological safety are also important steps toward ethical annotation pipelines (Spence et al., 2023).

8.2 Power Dynamics in Data Labour

Annotation work is often conducted through crowdworking platforms that rely on a globally distributed, low-cost labour force. These platforms are sometimes described as democratising access to work, but they often reflect significant power asymmetries between requesters and workers. Annotators frequently operate as anonymous contractors with no job security, limited bargaining power, and little visibility into how their work is used (Roberts, 2016). Compensation is usually task-based, which creates incentives to prioritise speed over accuracy.

This trade-off can result in lower-quality labels and increase the risk of bias in the final dataset (Snow et al., 2008b). Additionally, annotators rarely have channels for providing feedback about unclear instructions, ambiguous data, or annotation policies. As a result, a valuable feedback loop for improving annotation guidelines is often lost (Miceli and Posada, 2022).

These structural imbalances are not only ethical concerns; they also have technical implications. Poor working conditions can degrade data quality, obscure disagreement patterns, and exclude minority perspectives (Snow et al., 2008b). Creating fairer and more collaborative annotation systems, where annotators are treated as skilled contributors instead of disposable labour, can help ensure both ethical integrity and model reliability.

Ethical considerations must not be separated from methodological concerns. The conditions under which data are created shape their reliability, fairness, and downstream utility. Addressing annotation bias requires attention not only to technical design, but also to the social and economic contexts in which annotation work occurs.

9 Conclusion and Future Directions

Annotation bias remains a central challenge for multilingual and multimodal LLMs, shaping how models learn, generalise, and interact with diverse users. Mitigation requires both proactive measures (e.g., diverse annotators, refined guidelines) and reactive tools (e.g., bias detection, post-hoc adjustment), underpinned by ethical commitments to annotator well-being and fair labour.

Future work should prioritise community-driven annotation in marginalised contexts, culturally grounded benchmarks, and richer annotator metadata to improve fairness diagnostics, particularly in low-resource settings. LLMs themselves can assist as scalable annotation and bias-detection tools, but must be guided by real-world social and cultural contexts.

This paper contributes a typology of annotation bias, surveys detection methods across multilingual and cultural settings, and outlines mitigation strategies. We extend an ensemble-based method to multilingual settings to address label noise and inter-annotator disagreement, demonstrating its effectiveness on four socially sensitive tasks. Incorporating cultural awareness and accountability throughout the data pipeline will help NLP systems better reflect the diversity of human communication.

Acknowledgments

We thank the anonymous reviewers and program chairs for their insightful comments and constructive suggestions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ani Petrosyan. 2025. Most used languages online by share of websites 2025.

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

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68. PMID: 31313636.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Alanna Bergman and Cynda Hylton Rushton. 2023. Overcoming stigma: Asking for and receiving mental health support. *AACN advanced critical care*, 34(1):67–71.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

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.

Xia Cui, Terry Hanley, Muj Choudhury, and Tingting Mu. 2024. Data-driven or dataless? detecting indicators of mental health difficulties and negative life events in financial resilience using prompt-based learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 33–42.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

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.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 162–170, New York, NY, USA. Association for Computing Machinery.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models.

Boting Gao and Doug P VanderLaan. 2020. Cultural influences on perceptions of emotions depicted in emojis. *Cyberpsychology, Behavior, and Social Networking*, 23(8):567–570.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

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.

Hatice Gunes and Massimo Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345.

Kang He, Yinghan Long, and Kaushik Roy. 2024. Prompt-based bias calibration for better zero/few-shot learning of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12673–12691, Miami, Florida, USA. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Ziyi Huang, Nishanthi Rupika Abeynayake, and Xia Cui. 2025. Weak ensemble learning from multiple annotators for subjective text classification. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2025. The gaps between fine tuning and in-context learning in bias evaluation and debiasing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2758–2764, Abu Dhabi, UAE. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. The impact of debiasing on the performance of language models in downstream tasks is underestimated. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 29–36, Nusa Dua, Bali. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2022. Toward annotator group bias in crowdsourcing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1797–1806,

Dublin, Ireland. Association for Computational Linguistics.

Martin Lukac, Gulnaz Zhambulova, Kamila Abdiyeva, and Michael Lewis. 2023. Study on emotion recognition bias in different regional groups. *Scientific Reports*, 13(1):8414.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Milagros Miceli and Julian Posada. 2022. The data-production dispositif. *Proceedings of the ACM on human-computer interaction*, 6(CSCW2):1–37.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Saif M. Mohammad. 2016. 9 - sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Laurie Anne Pearlman and Karen W Saakvitne. 1995. *Trauma and the therapist: Countertransference and vicarious traumatization in psychotherapy with incest survivors.* WW Norton & Company.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sarah T. Roberts. 2016. *The Intersectional Internet: Race, Sex, Class and Culture Online*, chapter Commercial content moderation: Digital labourers'dirty work. Peter Lang Publishing.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Matthew Shardlow. 2022. Agree to disagree: Exploring subjectivity in lexical complexity. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 9–16, Marseille, France. European Language Resources Association.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nigel C Smeeton. 1985. Early history of the kappa statistic. *Biometrics*, 41:795.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008a. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008b. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. volume 72, page 1385–1470, El Segundo, CA, USA. AI Access Foundation.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. 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.

J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020a. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Haiping Zhang, Xingxing Zhou, Guoan Tang, Genlin Ji, Xueying Zhang, and Liyang Xiong. 2020b. Inference method for cultural diffusion patterns using a field model. *Transactions in GIS*, 24(6):1578–1601.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Yi Zhou, Jose Camacho-Collados, and Danushka Bollegala. 2023. A predictive factor analysis of social biases and task-performance in pretrained masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11082–11100, Singapore. Association for Computational Linguistics.

Supplementary Materials

A Relationships Among Bias Types, Detection, and Mitigation

This supplementary section details the observed relationships between annotation bias types and their associated detection and mitigation approaches, as illustrated in Figures 1 and 2. These relationships emerged from our analysis of current literature and empirical findings, rather than constituting a prescriptive framework.

A.1 Relationships Between Bias Types

Our analysis identifies three primary bias types that frequently interact in annotation processes: (1) **Instruction Bias**: Arising from task design choices, guidelines, and prompt formulations; (2) **Annotator Bias**: Stemming from individual predispositions and demographic characteristics; (3) **Contextual & Cultural Bias**: Emerging from cultural mismatches and linguistic norms. These bias types often co-occur and compound each other, particularly in multilingual settings where cultural context influences both task interpretation and annotation execution.

A.2 Correlations Between Detection and Mitigation Approaches

Figure 2 illustrates correlations observed between specific bias types and effective handling strategies.

Instruction bias correlations include detection through inter-annotator agreement metrics (Krippendorff, 2011) and model disagreement analysis (Geva et al., 2019), with mitigation through guideline refinement (Parmar et al., 2023) and in-context debiasing (Ganguli et al., 2023).

Annotator bias correlations involve detection via metadata analysis (Sap et al., 2019) with mitigation through diverse annotator recruitment (Bender and Friedman, 2018) and weak ensemble learning (Huang et al., 2025).

Contextual and cultural bias correlations include detection via multilingual divergence analysis (Huang and Yang, 2023) and cultural inference methods (Zhang et al., 2020a), with mitigation through culturally grounded taxonomies (Ponti et al., 2020) and post-hoc adjustments (Kaneko et al., 2023).

A.3 Emergent Cross-Connections

Our analysis reveals several emergent cross-connections where detection methods inform mitigation strategies: (1) Inter-annotator disagreement metrics often correlate with both instruction and annotator bias, suggesting applications for ensemble-based mitigation; (2) Cultural inference methods show relationships with both bias detection and the development of culturally-aware taxonomies; (3) Metadata analysis frequently informs both bias identification and targeted mitigation through annotator diversity initiatives. These relationships suggest that effective bias handling may benefit from considering detection methods not only as diagnostic tools but also as informants for mitigation strategy selection. However, these correlations should be interpreted as observed relationships rather than definitive prescriptions, as contextual factors may alter their applicability in specific settings.

B Benchmark Comparison using WEL on Multilingual LLMs

B.1 Data

We assess Weak Ensemble Learning (WEL) on the LeWiDi 2023 shared task datasets (Leonardelli et al., 2023), which are designed to evaluate generalisation across languages and domains. The benchmark consists of four corpora that vary in language, genre, and annotation protocol.

Three corpora (*ArMIS*, *HS-Brexit*, *MD-Agreement*) contain social media posts from *X*[1]. **ArMIS** comprises Arabic posts annotated for misogyny. **HS-Brexit** contains English posts labelled for Brexit-related hate speech. **MD-Agreement** consists of English posts annotated for offensiveness across multiple domains (e.g., BLM, elections, COVID-19); we disregard domain metadata and treat them uniformly.

The fourth corpus, *ConvAbuse*, contains English dialogues between users and conversational agents. Utterances are rated on a 5-point abuse scale ranging from -3 (highly abusive) to 1 (non-abusive). Following prior work, we reduce this to a binary classification task: *offensive* (< 0) vs. *non-offensive* (≥ 0). Multi-turn dialogues are flattened into single text sequences.

All datasets undergo standard preprocessing, including the removal of HTML tags, URLs, user mentions, punctuation, digits, non-ASCII characters, and redundant whitespace. Table 1 provides a summary of dataset statistics, including split sizes, annotator ranges, and total annotator counts.

[1] https://x.com/

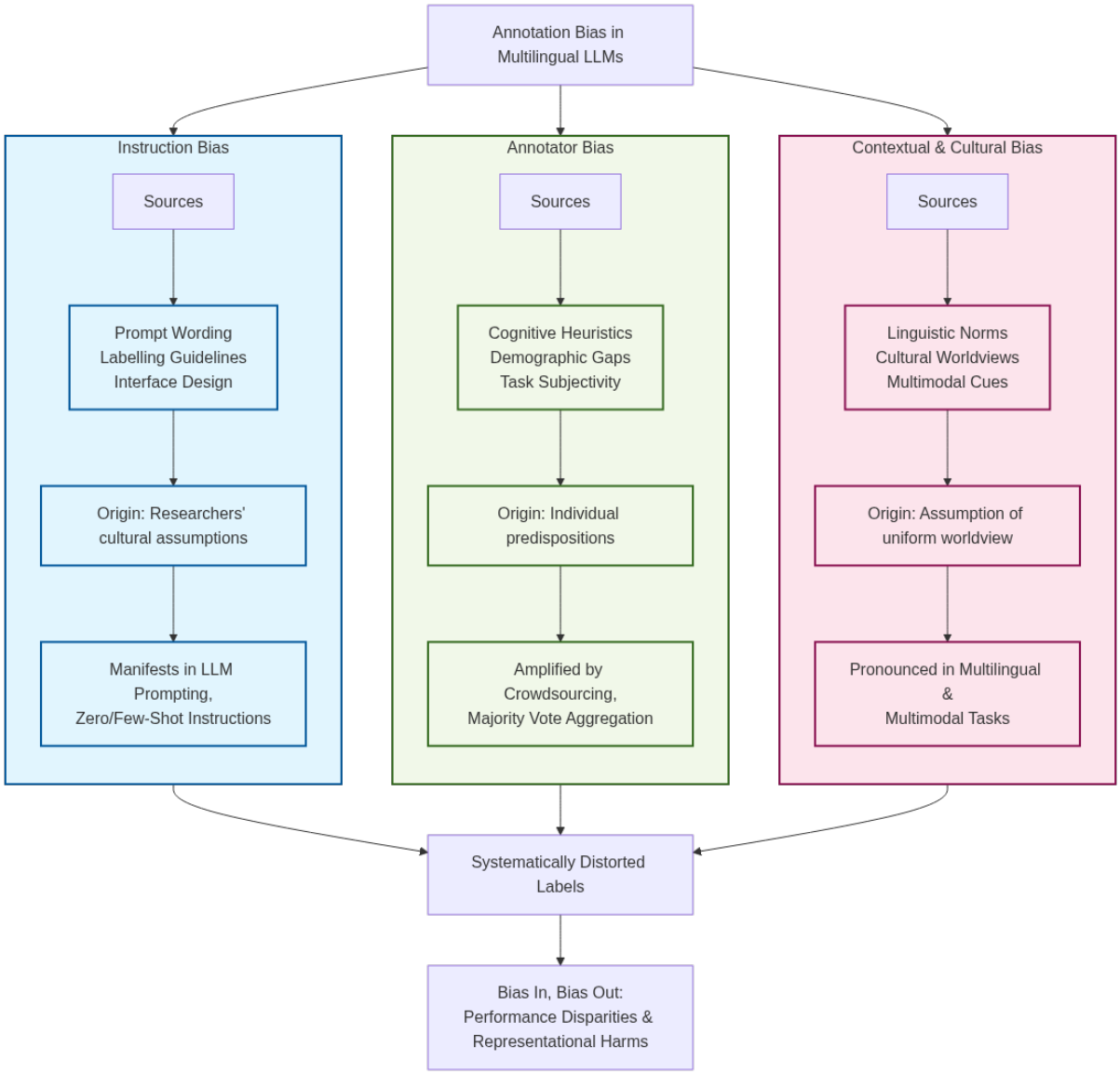

Figure 1: Taxonomy of annotation bias types observed in multilingual LLMs, showing three primary categories with distinct colouring: Instruction Bias, Annotator Bias, and Contextual & Cultural Bias.

B.2 Base LLM

To enable cross-linguistic generalisation in our ensemble-based framework, we adopt the multilingual BERT (`mBERT`) model (Devlin et al., 2019), more specifically, `bert-base-multilingual-uncased`[2], as the shared encoder for all weak learners in the WEL framework. This transformer-based model is pre-trained on 104 languages using a masked language modelling objective and retains casing information, making it well-suited for tasks with mixed scripts and morphologically rich languages.

In our setup, each weak predictor f_{θ_k} in the ensemble is instantiated by fine-tuning a separate copy of the multilingual BERT model on a label-variant dataset constructed via random sampling from annotator labels (as described in Section 7). Despite training on datasets in different languages and domains, all predictors share the same multilingual backbone, allowing for consistent representation across languages while preserving the benefits of ensemble diversity. This choice enables us to evaluate the robustness of WEL in a multilingual, multi-dataset context without requiring separate architectures per language.

[2] https://huggingface.co/google-bert/bert-base-multilingual-uncased

Table 1: Data statistics and metadata for the four textual datasets. #Train, #Dev, and #Test denote the number of instances in the training, development, and test splits. #TotalAnn indicates the total number of annotators, while #Ann represents the minimum and maximum number of annotators per instance. Contribution, Diversity, Language, and Genre provide further dataset details.

Dataset	#Train	#Dev	#Test	#TotalAnn	#Ann	Contribution	Diversity	Language	Genre
ArMIS	657	141	145	3	3	Fixed Annotators	Low	Arabic	Short Text
ConvAbuse	2398	812	840	8	2-7	Mixed Annotators	Low	English	Conversation
HS-Brexit	784	168	168	6	6	Fixed Annotators	Low	English	Short Text
MD-Agreement	6592	1104	3057	670	5	Mixed Annotators	High	English	Short Text

B.3 Results

Table 2 compares three models (CE-only (Uma et al., 2020), Top-5-Annotators (Xu et al., 2024), and WEL) across four datasets using F1 (higher better), cross-entropy (CE), and Manhattan distance (MD) (lower better). We perform a grid search over loss term coefficients in the objective function, each sampled from the range $[0, 0.001, 0.01, 0.1, 1]$, resulting in 1,295 unique combinations per dataset (excluding 0s for all). WEL consistently achieved the highest F1 scores and best calibration metrics (CE/MD) across three of four datasets, demonstrating its robustness for uncertainty-aware NLP.

Key observations emerge: (1) Performance varies substantially by domain, with *ConvAbuse* showing highest F1 scores but also extreme MD values for CE-only (4.81 vs. $<$1.0 elsewhere), indicating prediction instability that WEL addresses; (2) WEL's advantage in calibration metrics (CE/MD) exceeds its F1 improvements, highlighting its particular strength in uncertainty estimation; (3) Statistically significant improvements ($p < 0.05$) on *HS-Brexit* and *MD-Agreement* demonstrate WEL's robustness for hate speech and agreement tasks; (4) The *ArMIS* exception, where minimal gains occurred with only three annotators, establishes a practical boundary condition: WEL requires sufficient annotator diversity (likely $>$3) for effective ensemble learning. These results position WEL as particularly valuable for applications requiring reliable confidence estimates, while clearly defining its limitations in low-diversity annotation settings.

Table 2: Performance comparison across datasets and models. Best values are highlighted (F1: higher better; CE/MD: lower better). * indicates $p < 0.05$ significance.

Dataset	Metric	CE-only	Top-5-Ann	**WEL**
ArMIS	F1	0.6482	0.6552	0.6483
	CE	0.7019	0.6502	0.6596
	MD	0.7001	0.6443	0.6609
ConvAbuse	F1	0.8362	0.9310	0.9405*
	CE	0.9671	0.5651	0.5577*
	MD	4.8068	0.1648	0.1709
HS-Brexit	F1	0.7917	0.8929*	0.9167*
	CE	0.7652	0.6154*	0.5889*
	MD	0.7985	0.2394*	0.2585*
MD-Agreement	F1	0.7880	0.7808*	0.8214*
	CE	0.9948	0.6629*	0.6245*
	MD	1.7574	0.3995*	0.3632*

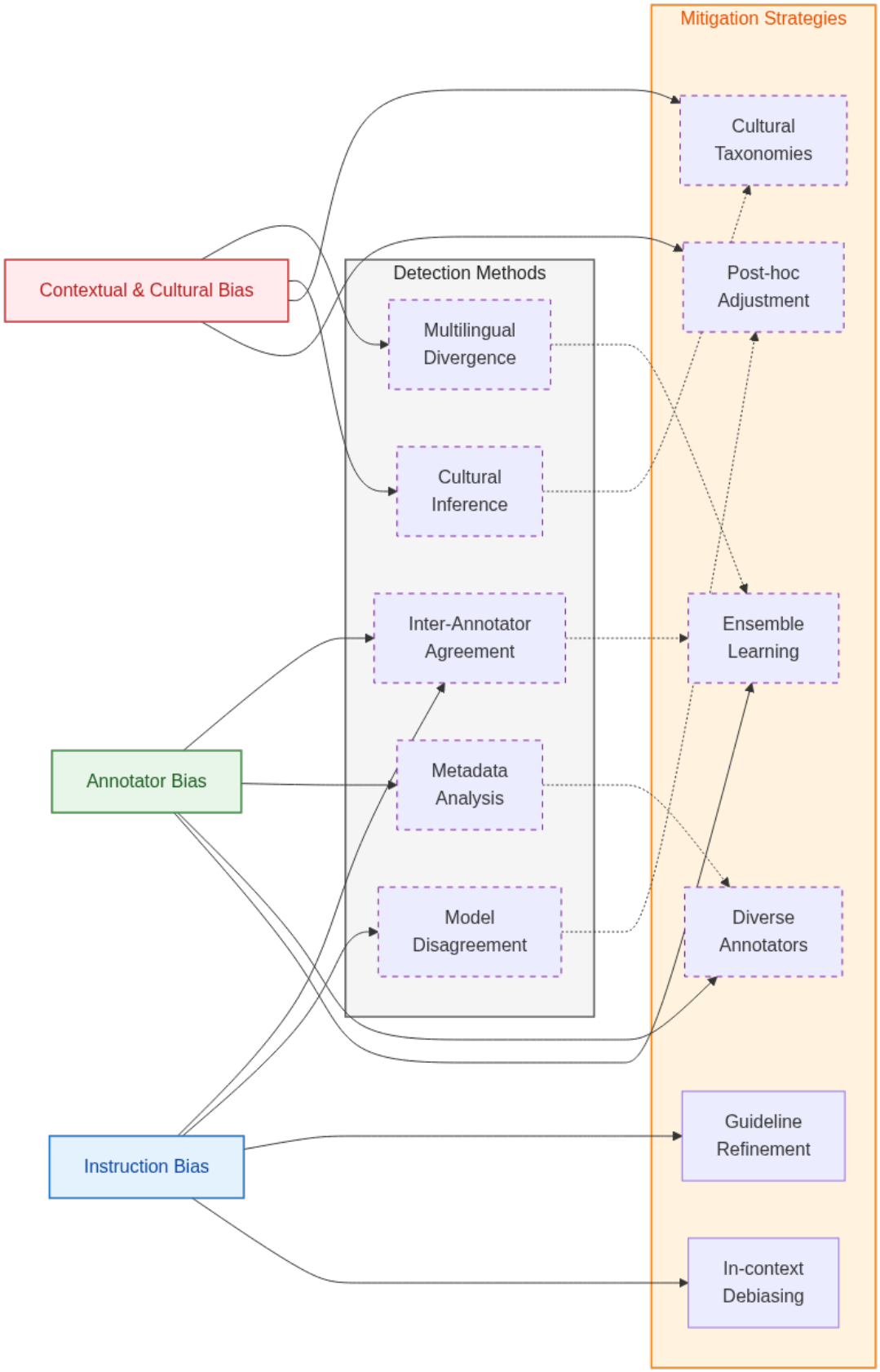

Figure 2: Relationships between annotation bias sources, detection methods, and mitigation strategies. Solid lines indicate primary correlations; dashed lines (purple) show secondary cross-connections.