Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Reclaiming the Loop: From the Consensus Trap to Pluralistic Data Annotation

Sheza Munir*
Computer Science, University of Toronto

Keywords: Data Annotation, Ground Truth, Epistemic Justice, Machine Learning

1. Introduction

Modern machine learning systems are built upon a "ground truth" paradigm that treats human disagreement as technical noise to be eliminated rather than a vital sociotechnical signal. This positivistic fallacy facilitates a "consensus trap" where standard practices like majority voting and model-mediated annotation systematically flatten the complexities of human judgment. By prioritizing procedural standardization (*niti*) over substantive justice (*nyaya*), current infrastructures enforce a logic of efficiency that erases marginalized perspectives. This research addresses two primary failures: the *allocation gap* (RQ1) — who annotates, and whether annotator–task alignment via lived experience predicts systematic label differences — and the *representation gap* (RQ2), concerning whether higher-alignment annotators produce qualitatively richer, contextually grounded rationales that are erased by majority-vote aggregation.

2. Proposed Solution and Approach

I propose the **Value-Driven Data Annotation (VDDA)** framework, a paradigm shift from extractive data labor toward *situated knowledge stewardship*. The core empirical contribution is an algorithm that quantifies annotator–task alignment via lived experience, recruits annotators differentiated by that alignment, and measures whether alignment predicts downstream labeling variance. Rather than eliminating disagreement, VDDA treats it as a high-fidelity signal — mapping the diversity of human experience instead of collapsing it to a single "right" answer.

3. Methodology: RQ1 – Who Annotates Matters

The VDDA framework operationalizes annotator–task alignment through a multi-component composite score:

$$S(a,t) \;=\; w_1 \cdot S_{\text{semantic}} \;+\; w_2 \cdot S_{\text{experience}} \;+\; w_3 \cdot S_{\text{perspective}} \;+\; w_4 \cdot S_{\text{reasoning}} \qquad (3.1)$$

where equal weights ($w_1 = w_2 = w_3 = w_4 = 0.25$) serve as the baseline, with ablation across weight configurations as a robustness check.

Lived experience is elicited via an AI-guided two-pass interface. In **Pass 1**, participants respond to a domain-specific vignette designed to surface interpretive differences without priming a label, including open-ended prompts: *What aspects of this situation stand out most to you, and why? Does this remind you of anything in your own life? What values guide how you evaluate whether something like this is harmful or supportive?* In **Pass 2**, participants articulate their epistemic standing: *"Based on what you have shared, what gives you standing to interpret content like this? What makes your perspective valuable or specific here?"* The interface synthesizes both passes into a 200–300 word positionality micronarrative per annotator, which feeds the four scoring components: semantic alignment

* sheza@cs.toronto.edu

(SBERT similarity to the task), relevant lived experience overlap, perspective alignment (LLM-extracted value stance similarity), and reasoning quality (LLM-scored depth and coherence), each normalized and combined for evaluation.

RQ1 tests whether S predicts systematic differences in annotation labels — specifically, whether high-alignment annotators identify harm, nuance, or ambiguity that low-alignment annotators systematically miss.

4. Methodology: RQ2 – Aggregating Across Reasoning Patterns

The post-annotation phase addresses the representation gap by replacing majority voting with rationale-aware aggregation. Rather than treating labels as votes to be counted, annotations are grouped by the semantic similarity of their accompanying justifications — identifying diverse but internally consistent reasoning patterns across the annotator pool [1, 2]. This shifts adjudication from procedural counting toward *perspectivist adjudication*: labels are evaluated based on the validity and sociocultural context of the reasoning behind them [3]. Minority signals are preserved as first-class data rather than averaged away. RQ2 tests whether annotators with higher alignment scores produce qualitatively richer, more contextually grounded rationales — and whether this holds across participation, advocacy, and expertise types of positionality.

5. Progress and Preliminary Impact

The foundational stage of this research involved a systematic literature review ($N = 346$) of papers published between 2020 and 2025 in seven target venues (ACL, AIES, CHI, CSCW, EAAMO, FAccT, and NeurIPS). This work deconstructs the positivistic fallacy of the ground truth paradigm and has been accepted for publication at FAccT 2026 [4].

- **Technical Validation:** Preliminary experiments indicate that individuals with direct lived experience provide up to 1.9× higher accuracy on domain-specific tasks compared to generic crowdworkers, validating that experiential depth is a functional necessity for data validity [5].
- **Outcome Shift:** Empirical evidence shows that altering the composition of a demographic jury can shift up to 14% of classification outcomes, demonstrating that "truth" is functionally a product of who is invited to the table [1].
- **Pilot Study:** A 10-participant pilot across 2–3 groups will validate three components prior to full deployment: micronarrative quality (richness sufficient for SBERT embedding), scoring function stability (does S vary sensibly across connection types, corroborated by two independent raters?), and interface usability (Streamlit elicitation and annotation flow tested end-to-end with prompts adjusted from participant feedback).

6. Conclusion and Future Work

The VDDA framework operationalizes the positionality collapse problem — identified in *The Consensus Trap* [4] — into a measurable, end-to-end annotation pipeline. Phases 3 and 4 (AI-guided elicitation and alignment scoring) are under active development and will be presented at the Symposium, where we invite community input on: weight configuration for w_1–w_4; validation thresholds for the scoring function; handling of sparse micronarratives from privacy-conscious participants; and mitigating LLM circularity in $S_{\text{reasoning}}$.

Future work will focus on dismantling geographic hegemony by integrating with informal local channels (e.g., WhatsApp guilds) to reach populations frequently dismissed as "hard-to-reach" [6], and on extending the pipeline across additional sensitive domains beyond the two datasets under current study.

References

[1] M. L. Gordon et al. “Jury Learning: Integrating Dissenting Voices into Machine Learning Models”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022.

[2] X. Geng and Q. Zhao. “Label Distribution Learning”. In: *CoRR* (2014). arXiv: 1408.6027.

[3] N. Sambasivan et al. “"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021.

[4] S. Munir, B. Mah, K. Kalsi, S. Kapania, J. Posada, E. Law, D. Wang, and S. I. Ahmed. “The Consensus Trap: Dissecting Subjectivity and the “Ground Truth” Illusion in Data Annotation”. In: *arXiv preprint arXiv:2602.11318* (2026). arXiv: 2602.11318 [cs.AI]. URL: https://arxiv.org/abs/2602.11318.

[5] S. Wallace et al. “Towards Fair and Equitable Incentives to Motivate Paid and Unpaid Crowd Contributions”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025.

[6] A. De et al. “Who Gets Heard? Calling Out the ’Hard-to-Reach’ Myth for Non-WEIRD Populations”. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2025.