Parameter-Driven Consensus-based Filtering Improves Collective Judgment Reliability in Crowdsourced Annotation

Jiyi Li (garfieldpigljy@gmail.com)[1]
[1]Hokkaido University

Abstract

Crowdsourced annotation can be viewed as a form of distributed human judgment in which individual reliability and task difficulty jointly shape collective decisions. Probabilistic aggregation models estimate latent annotator reliability and task difficulty, but are typically used only to weight judgments or infer labels, not to guide selective dataset refinement. We propose a replicated-dataset consensus filtering framework that improves collective judgment reliability by removing unreliable annotators and difficult tasks based on latent reliability-difficulty parameters. Instead of relying on a single dataset estimate, the method constructs multiple replicated datasets by small random label removal, re-estimates latent parameters on each replicated dataset, and removes components that are consistently selected across replicated datasets. Filtering is applied iteratively with entropy-based stopping conditions. The framework is model-agnostic and can be combined with any probabilistic aggregation model that estimates reliability and difficulty parameters. Using a standard reliability-difficulty model instantiation, experiments on multiple crowdsourced datasets show that parameter-driven consensus filtering improves aggregation accuracy and removes incorrect tasks more selectively than random removal. These results support a cognitively grounded, parameter-based approach to improving collective judgment reliability in distributed annotation settings.

Keywords: collective judgment; crowdsourcing; judgment reliability; parameter-based filtering; cognitive uncertainty

Introduction

Many cognitive science studies rely on judgments collected from multiple human participants, including perception, language interpretation, categorization, and reasoning tasks. Large-scale crowdsourced annotation provides a practical testbed for studying such distributed human judgment under uncertainty, while also serving as a common way to construct labeled datasets for AI and behavioral research. From a cognitive perspective, crowdsourced annotation can be viewed as distributed human judgment (Surowiecki, 2005), in which individuals independently produce categorical decisions under varying levels of uncertainty (Tversky & Kahneman, 1974). This view aligns with recent cognitive research on group estimation and uncertainty reduction in collective judgment tasks (Salikutluk & Jäkel, 2025).

In such settings, data quality is not only a technical annotation problem but also a question of how heterogeneous judgments should be combined. Participants differ in expertise, attention, strategy, reliability, and interpretation, while stimuli differ in ambiguity and cognitive demand. Disagreement among annotators is therefore not merely random noise, but may reflect latent cognitive heterogeneity in both judges and tasks.

The most common aggregation strategy for crowd judgments is majority voting, which treats all annotators as equally reliable. While simple and often effective, equal-weight aggregation assumes uniform judgment quality and independent errors. In practice, however, annotators vary substantially in reliability, and tasks vary in difficulty and ambiguity. When cognitive load is high or task interpretation is uncertain, disagreement increases in systematic rather than random ways. Treating disagreement purely as noise can therefore obscure meaningful reliability structure in collective judgment. Recent reviews of collective intelligence highlight how cognitive diversity and interactive deliberation can reduce uncertainty and improve aggregate decisions (Cui & Yasseri, 2024).

To address annotator variability, probabilistic aggregation models explicitly estimate worker reliability and latent true labels (Dawid & Skene, 1979; Karger et al., 2011; Liu et al., 2012; Venanzi et al., 2014; D. Zhou et al., 2012). A particularly influential model, GLAD (Whitehill et al., 2009), jointly estimates annotator ability and task difficulty, modeling correctness as a function of both factors. These parameters are cognitively interpretable as individual reliability and task-level difficulty. However, prior work primarily uses such parameters to weight judgments or infer labels, rather than to guide selective dataset refinement.

In many applications, the goal is not only to infer labels but also to improve the quality of the labeled dataset itself by identifying unreliable annotators and difficult tasks. We therefore study parameter-driven filtering for collective annotation quality control. We propose a replicated-dataset consensus filtering framework that operates on latent reliability and difficulty parameters estimated by a probabilistic aggregation model. Instead of relying on a single dataset estimate, the method constructs multiple replicated datasets by small random label removal, re-estimates latent parameters on each replicated dataset, and removes annotators and tasks that are consistently selected across replicated datasets. The filtering process is applied iteratively with entropy-based stopping conditions. Our contribution is not a new reliability model, but a new consensus-based use of latent reliability and difficulty parameters for selective component removal.

Empirically, using a standard reliability-difficulty model in-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

stantiation (GLAD), we show that parameter-driven consensus filtering improves aggregation accuracy and removes incorrect tasks more selectively than random removal across multiple crowdsourced datasets. These results support using latent reliability and difficulty parameters not only for label inference but also for controlled dataset refinement in distributed judgment settings.

Existing reliability-aware aggregation methods mainly treat latent reliability and difficulty as inference weights: unreliable judgments are down-weighted, but the annotation pool itself remains unchanged. In contrast, our framework uses these parameters as reproducible selection signals for changing the composition of the dataset. The key difference is therefore not the parameter estimator itself, but the decision layer built on top of it: replicated perturbation, cross-replication consensus, and conservative component removal.

The contributions of this paper are as follows. First, we introduce a replicated-dataset consensus filtering framework for collective annotation quality control based on latent reliability and difficulty parameters. Second, we present a model-agnostic filtering pipeline that combines replicated datasets, consensus-based component removal, and entropy-based stopping. Third, we empirically demonstrate that parameter-driven consensus filtering improves aggregation accuracy and selectively removes incorrect tasks under established experimental settings.

Related Work

Quality control in crowdsourcing has been widely studied as a problem of aggregating heterogeneous human judgments under uncertainty. Majority voting (Snow et al., 2008) is one of the simplest and most widely used aggregation strategies. While often effective, it assigns equal weight to all workers and assumes uniform reliability. In realistic annotation settings, however, workers differ substantially in reliability, diligence, and interpretation (Hong & Page, 2004), making equal-weight aggregation cognitively unrealistic and potentially unreliable at the group level.

To address this limitation, a large body of work has developed probabilistic aggregation models that explicitly estimate worker reliability and latent true labels. A foundational approach is the EM-based observer error model of Dawid and Skene (1979) (also known as Dawid-Skene), which estimates worker-specific error rates and true labels jointly. Together, these probabilistic aggregation models treat disagreement as structured signal rather than random noise and form the statistical foundation of reliability-aware crowd aggregation (Karger et al., 2011; Liu et al., 2012; Venanzi et al., 2014; D. Zhou et al., 2012). A particularly influential line of work jointly models worker ability and task difficulty (Bachrach et al., 2012; Welinder et al., 2010). The GLAD model (Whitehill et al., 2009) parameterizes the probability of correct judgment as a function of annotator ability and item difficulty.

More recent approaches explore alternative machine learning formulations for crowd quality modeling, including tensor completion, representation learning, graph-based aggregation, and so on (Kawase et al., 2019; D. Li & de Rijke, 2023; J. Li, 2023, 2024a, 2025; J. Li & Kashima, 2017; J. Li et al., 2017, 2018a, 2020; Yang et al., 2024; Yin et al., 2017; Y. Zhou & He, 2016). While these approaches improve aggregation performance, they are still mainly designed for label inference and confidence estimation rather than parameter-driven component filtering. We mainly focus on categorical human judgments in this work; there are also aggregation-based approaches for other types of human judgments, such as pairwise, triplet, and text judgments (Baba et al., 2020; J. Li, 2020, 2022, 2024b; J. Li & Fukumoto, 2019; J. Li et al., 2018b, 2021; Zhang et al., 2022; Zuo et al., 2020). Besides aggregation-based approaches, another line of work directly trains classification models using noisy crowd labels (J. Li et al., 2022; Lu et al., 2023; Rodrigues & Pereira, 2018), focusing on robust learning under label noise rather than explicit worker-task reliability modeling.

In summary, prior work has developed strong models for estimating reliability, difficulty, and latent labels. However, these estimates are usually used to improve inference within a fixed annotation pool. Less attention has been paid to how stable parameter estimates can be used to decide which annotators or tasks should be removed from the dataset. Our work addresses this gap by introducing replicated-dataset consensus filtering as a conservative decision layer on top of existing aggregation models: latent worker reliability and task difficulty parameters are used as selection signals, and only components repeatedly selected across replicated datasets are removed.

Our Approach

Overview

We propose a replicated-dataset consensus filtering framework for improving collective judgment reliability in crowdsourced annotation. The framework treats crowd labeling as a distributed judgment process and uses latent annotator reliability and task difficulty parameters estimated by a probabilistic aggregation model.

The method consists of two stages: (1) estimation of latent reliability and difficulty parameters from observed annotations, and (2) replicated-dataset consensus filtering. Multiple replicated datasets are constructed by randomly removing a small proportion of labels. Parameter estimation is performed on each replicated dataset, and annotators or tasks that are consistently selected for removal across replicated datasets are removed by consensus. The filtering process is applied iteratively with entropy-based stopping conditions. An overview of the replicated-dataset consensus filtering framework is shown in Figure 1.

Our contribution is not a new reliability model, but a model-agnostic filtering procedure that uses latent reliability and difficulty parameters for selective dataset refinement. The novelty lies in the replicated-dataset consensus removal mechanism, which converts latent reliability-difficulty parameters from weighting signals into reproducible component selection

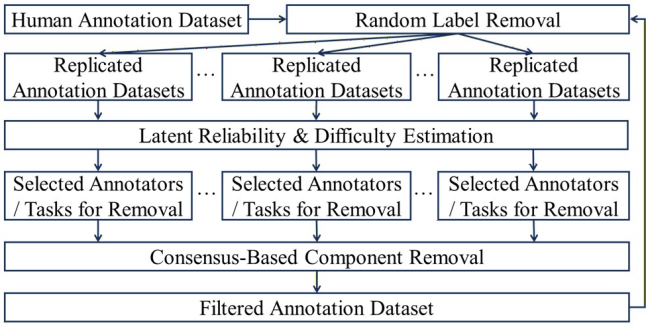

Figure 1: Replicated-dataset consensus filtering pipeline. The original annotation matrix is provided by human annotators, while replicated dataset generation, parameter estimation, consensus selection, and stopping decisions are performed automatically by the filtering framework.

criteria for dataset refinement. Any probabilistic model that estimates such parameters can be integrated into this framework. In our experiments, we instantiate the parameter estimation stage using the GLAD model, but the filtering pipeline itself does not depend on a specific estimator.

Conceptual Role of Parameter-Driven Filtering The proposed framework separates two functions that are often conflated in probabilistic aggregation: label inference and component selection. Traditional aggregation methods use latent reliability and difficulty parameters primarily to weight judgments during inference. In contrast, our framework uses these parameters as selection signals for dataset refinement. This distinction is important conceptually: weighting adjusts influence, whereas filtering alters the composition of the judgment pool. From a systems perspective, weighting assumes that all components should remain in the aggregation process but contribute unequally, while filtering assumes that some components are structurally harmful and should be excluded. The replicated-dataset consensus mechanism operationalizes this distinction by requiring reproducible selection across multiple parameter estimation runs before removal is applied.

Latent Reliability and Difficulty Estimation

Let i index annotators and j index tasks. Let L_{ij} denote the label provided by annotator i on task j, and let Z_j denote the latent true label. We assume a probabilistic aggregation model that introduces latent annotator reliability parameters and latent task difficulty parameters, inferred from observed labels. These parameters characterize systematic differences in annotator judgment quality and task-level difficulty. They are not directly observed but are estimated statistically from the annotation matrix using an EM-style or Bayesian inference procedure, depending on the specific model.

One common instantiation models the probability that annotator i produces a correct judgment on task j as a logistic function of annotator reliability and task easiness:

$$P(L_{ij} = Z_j \mid \alpha_i, \beta_j) = \frac{1}{1 + \exp(-\alpha_i \beta_j)} \tag{1}$$

where α_i represents annotator reliability and β_j represents task easiness (inverse difficulty). Higher α_i corresponds to more reliable judgment behavior, while lower β_j corresponds to higher task difficulty or ambiguity.

Given all labels for task j, denoted $\mathbf{L}_j$, the posterior probability of the true label is estimated as:

$$P(Z_j \mid \mathbf{L}_j, \alpha, \beta) \propto P(Z_j) \prod_i P(L_{ij} \mid Z_j, \alpha_i, \beta_j) \tag{2}$$

In standard aggregation settings, these parameters are used to infer labels or weight annotator votes. In our framework, they are additionally used to guide component removal.

Latent reliability and difficulty parameters are statistical constructs derived from annotation patterns rather than direct psychological measurements. However, they admit cognitively meaningful interpretations. Reliability parameters capture consistent annotator-level agreement patterns relative to inferred truth, while difficulty parameters capture task-level disagreement structure.

In this work, we treat these parameters as behavioral signals rather than psychological measurements. Their operational role is diagnostic rather than explanatory: they are used to guide selective component removal when supported by cross-replicated estimation agreement.

Replicated Datasets by Random Label Removal

Parameter estimates derived from a single dataset may be sensitive to label sparsity and uneven annotation patterns. To obtain more robust removal decisions, we construct multiple replicated datasets derived from the original annotation set.

Let x be the total number of labels. We generate k replicated datasets by randomly removing a small portion of labels from the original dataset. In our setting, each replicated dataset is constructed by removing $n = x/100$ (e.g., 1%) labels uniformly at random. This preserves the overall annotator-task structure while introducing small variations in observed evidence.

Latent reliability and difficulty parameters are then re-estimated independently for each replicated dataset using the chosen probabilistic aggregation model. Each run produces a candidate set of annotators and tasks selected for removal based on parameter criteria (e.g., low reliability or high difficulty).

Consensus-Based Component Removal

Each replicated dataset produces a set of annotators and tasks identified for removal according to parameter-based rules. Rather than removing components based on a single dataset estimate, we apply a consensus rule across replicated datasets.

Let $U^{(r)}$ denote the set of components selected for removal in replicated dataset r. The final removal set is defined as the intersection across replicated datasets:

$$U^* = \bigcap_{r=1}^{k} U^{(r)} \tag{3}$$

Only annotators and tasks contained in U^* are removed from the original dataset. This consensus-based removal rule ensures that components are filtered only when they are repeatedly identified across replicated datasets, reducing sensitivity to small label variations in any single run. We use strict intersection rather than frequency thresholds to enforce conservative selection and minimize false-positive removals caused by parameter estimation variance across replicated datasets.

After removal, the probabilistic aggregation model is reapplied to the filtered dataset to obtain updated parameter estimates and inferred labels.

Iterative Filtering and Entropy-Based Stopping

The consensus-based removal procedure is applied iteratively. After each removal step, parameter estimation and label inference are recomputed on the filtered dataset, replicated datasets are regenerated, and consensus removal is repeated.

The process stops when one of the following conditions is satisfied: (1) no annotators or tasks are commonly selected across replicated datasets; (2) the number of selected tasks exceeds a predefined upper bound to preserve dataset coverage; (3) the entropy of the inferred label distribution becomes zero, indicating deterministic label assignments by the aggregation model. The entropy-based stopping condition reflects convergence of label inference confidence and serves as a practical termination criterion.

Interpretation

In this framework, latent reliability parameters capture individual differences in judgment quality, and latent difficulty parameters capture task-level ambiguity and cognitive load. Replicated datasets created by small random label removal provide multiple views of the same annotation structure. Consensus-based component removal selects annotators and tasks that are consistently identified under these replicated views.

The overall method combines probabilistic parameter estimation, replicated datasets, consensus removal, and entropy-based stopping into a model-agnostic filtering pipeline for improving collective label quality.

Filtering Design Rationale

The filtering design follows three principles: small perturbation, cross-replication agreement, and conservative removal. First, replicated datasets are generated through small random label removal to introduce controlled variation without changing the overall annotator-task topology. This allows testing whether parameter-based selection decisions are stable under minor evidence changes.

Second, consensus-based selection requires agreement across replicated datasets. This reduces sensitivity to estimation noise and prevents removal decisions driven by single-run parameter fluctuations. Only components repeatedly selected under parameter-based criteria are removed.

Third, entropy-based stopping constraints ensure that filtering does not over-prune the dataset. Removal stops when inferred label distributions reach high confidence or when predefined removal limits are reached. Together, these design choices prioritize reproducibility and conservativeness over aggressive pruning.

Table 1: Crowdsourced annotation datasets.

Dataset	#Tasks	#Annotators	#Labels
Duck	108	39	4,212
Weather	300	110	6,000
Temp	462	76	4,620
RTE	800	164	8,000

Experiments

Datasets

We evaluate the proposed replicated-dataset consensus filtering framework on four crowdsourced annotation datasets with redundant labels and gold-standard answers. These datasets are widely used benchmarks for studying aggregation under annotator reliability variation and task difficulty differences. They cover perceptual, sentiment, temporal reasoning, and textual inference tasks, providing diversity in annotation difficulty and judgment structure. The datasets include **Duck** (Welinder et al., 2010): identify whether the image contains a duck; **Weather** (Venanzi et al., 2015): judge the sentiment of a tweet discussing the weather; **Temp** (Temporal Ordering) (Snow et al., 2008): identify whether one event happens before another event in context; **RTE** (Recognizing Textual Entailment) (Snow et al., 2008): judge whether a hypothesis sentence is implied by a given text.

Each dataset contains multiple annotations per task from different annotators. This redundancy enables estimation of latent annotator reliability and task difficulty parameters using a probabilistic aggregation model. In our experiments, this model is instantiated using GLAD, but the filtering framework itself is model-agnostic. GLAD is used as a representative reliability-difficulty estimator; the filtering mechanism operates on estimated parameters and does not depend on model-specific structure.

Evaluation Method

We use two evaluation metrics in the experiments: (1) **aggregation accuracy**, defined as agreement between inferred labels and ground-truth labels; and (2) **incorrect-task removal ratio**, defined as the proportion of removed tasks that are actually incorrectly labeled, compared against random task removal at the same removal rate.

Accuracy measures overall aggregation quality, while incorrect-task removal ratio measures how selectively the filtering procedure targets error-prone tasks. These two metrics jointly evaluate both aggregation quality and removal selec-

tivity, corresponding to label correctness and component selection precision.

For each benchmark dataset, aggregation accuracy was computed against the gold-standard labels provided with the dataset. The same gold labels were used to evaluate majority voting, the probabilistic model without filtering, and the proposed consensus-filtering method. For the incorrect-task removal ratio, a removed task was counted as incorrect when the model-inferred label on the unfiltered dataset did not match the gold-standard label.

Experimental Procedure

For each dataset, we estimate latent annotator reliability and task difficulty parameters using a probabilistic aggregation model, i.e., GLAD in our experiments. These parameters are used to select unreliable annotators and difficult tasks for removal.

To make removal decisions robust, we construct replicated datasets by random label removal. Let x be the total number of labels. We generate $k = 10$ replicated datasets, each formed by randomly removing $n = x/100$ (1%) labels from the original dataset. Parameter estimation is performed independently on each replicated dataset. The small removal rate introduces controlled perturbations while preserving the overall annotation structure, allowing removal decisions to be tested for cross-replication consistency without altering dataset topology.

Each replicated dataset produces a set of annotators and tasks selected for removal according to parameter-based criteria. Only components selected across replicated datasets are included in the consensus removal set. These annotators and tasks are removed from the original dataset, and aggregation is recomputed on the filtered dataset.

Filtering is applied iteratively until one of the predefined stopping conditions is met, including convergence of inferred labels (zero entropy) or reaching an upper bound on the number of removed tasks.

Several design choices aim to balance robustness and minimal intervention. The replicated-dataset strategy introduces small perturbations to test reproducibility of parameter-based selection without materially altering dataset structure. A 1% label removal rate was chosen to maintain annotation density while enabling cross-run variation.

Consensus-based intersection is used instead of union to enforce conservative selection. This reduces false-positive removals caused by estimation noise. Random removal is used as a baseline because it is structure-agnostic and provides a direct test of whether parameter-driven removal is selectively targeting problematic components rather than removing items arbitrarily.

Experimental Results

Table 2 reports aggregation accuracy for majority voting, probabilistic aggregation without filtering, and probabilistic aggregation with parameter-driven consensus filtering.

Table 2: Aggregation accuracy (%) across methods.

Dataset	Consensus + Model	Model Only	Majority Voting
Duck	**77.10**	72.22	75.93
Weather	**91.14**	85.67	84.67
Temp	**95.43**	93.51	93.94
RTE	**95.71**	92.50	91.88

Table 3: Incorrect-task ratio among removed tasks (%).

Dataset	Consensus Filtering	Random Removal
Duck	**100.00**	27.78
Weather	**43.75**	14.23
Temp	**60.00**	6.94
RTE	**39.13**	7.50

Across all datasets, consensus filtering improves accuracy over both majority voting and probabilistic aggregation alone.

Table 3 shows the proportion of removed tasks that are actually incorrect, compared with random removal at the same removal rate. Random removal serves as a structure-agnostic baseline to test whether parameter-based removal is selective rather than arbitrary. Consensus filtering removes substantially higher proportions of incorrect tasks than random selection, indicating that parameter-driven removal is selective rather than arbitrary.

Together with the removal-selectivity results, these findings indicate that latent reliability and difficulty parameters provide effective signals for selective dataset refinement under consensus-based filtering.

Beyond overall accuracy gains, the removal selectivity results indicate that parameter-driven consensus filtering is not merely reducing dataset size but preferentially excluding components associated with incorrect labels. The comparison with random removal under equal removal budgets shows that filtering decisions are structured rather than arbitrary. This supports interpreting latent parameter estimates as useful diagnostic signals for dataset refinement. The selective removal effect suggests that latent reliability and difficulty parameters encode structured judgment signals beyond simple noise levels. Tasks that are repeatedly removed under consensus filtering tend to exhibit higher disagreement dispersion and lower cross-annotator consistency, while retained tasks tend to show more coherent judgment structure. This pattern indicates that parameter-driven consensus filtering functions as a structure-aware selection mechanism that preserves judgment coherence while reducing unreliable variation, rather than acting as a uniform data reduction procedure.

Discussion

Collective Annotation as Distributed Cognitive Judgment

Our results support interpreting crowdsourced annotation as a distributed cognitive judgment system shaped by heterogeneous annotator reliability and task-dependent difficulty. Annotator disagreement should therefore not be treated purely

as random noise. Instead, disagreement patterns reflect structured variation that can be captured by latent reliability and difficulty parameters. Parameter-driven consensus filtering improves aggregation accuracy because it uses this structure to guide selective component removal rather than uniformly averaging across all judgments. These interpretations are consistent with cognitive reliability and task load theories, though the parameters themselves are statistically inferred rather than directly measured.

From a cognitive perspective, collective labeling is a multi-agent judgment process in which contributors differ in reliability and tasks differ in cognitive demand. Our findings suggest that collective performance depends not only on averaging more judgments, but also on regulating which judgment sources and tasks are retained. In this sense, parameter-based consensus removal provides a computational account of reliability regulation in distributed judgment systems by identifying unreliable annotators and consistently problematic tasks that degrade group-level inference.

Reliability and Difficulty Parameters as Behavioral Signals

Probabilistic aggregation models estimate latent annotator reliability and task difficulty parameters (Dawid & Skene, 1979; Whitehill et al., 2009). Our experiments support interpreting these parameters as behavioral signals of judgment quality and cognitive load. Annotators with low inferred reliability and tasks with high inferred difficulty are more frequently selected under consensus-based removal criteria and are associated with lower aggregation reliability.

Importantly, removal decisions are based on agreement across replicated datasets rather than single-run estimates. This consensus requirement reduces sensitivity to estimation noise and supports more conservative, repeatable component selection. In this sense, replicated-dataset agreement functions as a reproducibility constraint on parameter-based filtering. Empirically, this parameter-driven selection removes incorrect tasks at substantially higher rates than random removal under the same removal budget.

Implications for Annotation Quality Control

These findings suggest that reliability-difficulty parameter estimation can support not only label inference but also principled dataset refinement. Annotation quality control can use parameter-driven consensus removal to selectively exclude unreliable annotators and difficult tasks, rather than relying only on vote weighting or post-hoc accuracy checks.

More broadly, parameter-based consensus filtering provides a cognitively interpretable and model-compatible mechanism for reliability regulation in distributed judgment systems. It suggests that collective annotation quality can be improved not only by collecting more judgments, but also by regulating which judgment sources and tasks are retained.

Risk of Repeated Filtering

A potential risk of repeated parameter-driven filtering is that early estimation errors may be amplified across iterations, leading to over-pruning or reinforcing biases in the initial model estimates. Our framework mitigates this risk through small perturbations, consensus-based removal, and conservative stopping conditions. Replicated datasets preserve the original annotation topology, and components are removed only when consistently selected across replicated datasets, reducing dependence on any single estimate. Filtering also stops when no consensus removal exists, when a removal upper bound is reached, or when inferred labels become deterministic. Nevertheless, model-induced bias remains a limitation, especially when the initial aggregation model is misspecified or when minority but valid interpretations are treated as errors.

Random Guessing and Adversarial Annotation

The current framework identifies annotators who are unreliable with respect to the inferred consensus, but it does not fully diagnose the source of unreliability. Random guessing would likely produce near-chance agreement and noisy reliability estimates, whereas adversarial annotation may produce systematic disagreement or label-flipped patterns, especially in binary tasks. However, the GLAD-style instantiation used here is not designed to explicitly distinguish these behaviors. Doing so would require extensions such as signed reliability parameters, worker-specific confusion matrices, adversarial-aware aggregation models, or validation questions with known answers. Thus, the present method should be viewed as a conservative filtering mechanism rather than a complete diagnosis of annotator intent.

Conclusion

We presented a parameter-driven consensus filtering framework for improving collective judgment reliability in crowdsourced annotation. The approach uses latent annotator reliability and task difficulty parameters estimated by a probabilistic aggregation model to guide selective component removal through replicated datasets and consensus-based selection. Unlike prior work that uses parameters primarily for weighting, we use them for reproducible component selection. Experiments across multiple datasets show that parameter-based consensus filtering improves aggregation accuracy and removes incorrect tasks more selectively than random removal under the same removal budget. The improvement comes from structured component selection rather than more complex aggregation models.

From a cognitive perspective, these results support viewing crowdsourced annotation as a distributed judgment process shaped by heterogeneous reliability and task-dependent cognitive demand. Future work should therefore evaluate adversarial-aware parameterizations, signed reliability models, and worker-specific confusion-matrix models to distinguish different forms of unreliable annotation behavior.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K28092. Thanks Ikuno Nishikawa for his work on part of this topic in his undergraduate thesis.

References

Baba, Y., Li, J., & Kashima, H. (2020). Crowdea: Multi-view idea prioritization with crowds. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, *8*(1), 23–32.

Bachrach, Y., Minka, T., Guiver, J., & Graepel, T. (2012). How to grade a test without knowing the answers: A bayesian graphical model for adaptive crowdsourcing and aptitude testing. *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, 819–826.

Cui, H., & Yasseri, T. (2024). Ai-enhanced collective intelligence. *Patterns*, *5*(11), 101074.

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

Hong, L., & Page, S. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *PNAS*.

Karger, D. R., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. *Advances in Neural Information Processing Systems (NIPS)*, 1953–1961.

Kawase, Y., Kuroki, Y., & Miyauchi, A. (2019). Graph mining meets crowdsourcing: Extracting experts for answer aggregation. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 1272–1279.

Li, D., & de Rijke, M. (2023). Extending label aggregation models with a gaussian process to denoise crowdsourcing labels. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 729–738.

Li, J., Sun, H., & Li, J. (2022). Beyond confusion matrix: Learning from multiple annotators with awareness of instance features. *Mach. Learn.*, *112*(3), 1053–1075.

Li, J. (2020). Crowdsourced text sequence aggregation based on hybrid reliability and representation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1761–1764.

Li, J. (2022). Context-based collective preference aggregation for prioritizing crowd opinions in social decision-making. *Proceedings of the ACM Web Conference 2022 (WWW)*, 2657–2667.

Li, J. (2023). Learning representations for sparse crowd answers. *Proceedings of the 30th International Conference on Neural Information Processing (ICONIP)*, 468–480.

Li, J. (2024a). A comparative study on annotation quality of crowdsourcing and llm via label aggregation. *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6525–6529.

Li, J. (2024b). Human-llm hybrid text answer aggregation for crowd annotations. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 15609–15622.

Li, J. (2025). Label aggregation of composite crowd tasks by worker ability constraint satisfaction. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*.

Li, J., Baba, Y., & Kashima, H. (2017). Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, 1069–1078.

Li, J., Baba, Y., & Kashima, H. (2018a). Incorporating worker similarity for label aggregation in crowdsourcing. *Proceedings of the 27th International Conference on Artificial Neural Networks (ICANN)*, 596–606.

Li, J., Baba, Y., & Kashima, H. (2018b). Simultaneous clustering and ranking from pairwise comparisons. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 1554–1560.

Li, J., Endo, L. R., & Kashima, H. (2021). Label aggregation for crowdsourced triplet similarity comparisons. *Proceedings of the 28th International Conference on Neural Information Processing (ICONIP)*, 176–185.

Li, J., & Fukumoto, F. (2019). A dataset of crowdsourced word sequences: Collections and answer aggregation for ground truth creation. *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP (AnnoNLP)*, 24–28.

Li, J., & Kashima, H. (2017). Iterative reduction worker filtering for crowdsourced label aggregation. *Proceedings of the 18th International Conference on Web Information Systems Engineering (WISE)*, 46–54.

Li, J., Kawase, Y., Baba, Y., & Kashima, H. (2020). Performance as a constraint: An improved wisdom of crowds using performance regularization. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 1534–1541.

Liu, Q., Peng, J., & Ihler, A. (2012). Variational inference for crowdsourcing. *Advances in Neural Information Processing Systems (NIPS)*, 692–700.

Lu, X., Li, J., Takeuchi, K., & Kashima, H. (2023). Multiview representation learning from crowdsourced triplet comparisons. *Proceedings of the ACM Web Conference 2023 (WWW)*, 3827–3836.

Rodrigues, F., & Pereira, F. C. (2018). Deep learning from crowds. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

Salikutluk, V., & Jäkel, F. (2025). Deliberation in guesstimation. *Cognitive Science*, *49*(8), e70090.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the*

Conference on Empirical Methods in Natural Language Processing (EMNLP), 254–263.

Surowiecki, J. (2005). *The wisdom of crowds*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., & Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, 155–164.

Venanzi, M., Teacy, W., Rogers, A., & Jennings, N. R. (2015). Weather sentiment - amazon mechanical turk dataset.

Wauthier, F. L., & Jordan, M. (2011). Bayesian bias mitigation for crowdsourcing. *Advances in Neural Information Processing Systems (NIPS)*, *24*.

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The multidimensional wisdom of crowds. *NIPS*, 2424–2432.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS)*, 2035–2043.

Yang, Y., Zhao, Z.-Q., Wu, G., Zhuo, X., Liu, Q., Bai, Q., & Li, W. (2024). A lightweight, effective, and efficient model for label aggregation in crowdsourcing. *ACM Trans. Knowl. Discov. Data*, *18*(4).

Yin, L., Han, J., Zhang, W., & Yu, Y. (2017). Aggregating crowd wisdoms with label-aware autoencoders. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 1325–1331.

Zhang, G., Li, J., & Kashima, H. (2022). Improving pairwise rank aggregation via querying for rank difference. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–9.

Zhou, D., Platt, J. C., Basu, S., & Mao, Y. (2012). Learning from the wisdom of crowds by minimax entropy. *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 2195–2203.

Zhou, Y., & He, J. (2016). Crowdsourcing via tensor augmentation and completion. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 2435–2441.

Zuo, X., Li, J., Zhou, Q., Li, J., & Mao, X. (2020). Affecti: A game for diverse, reliable, and efficient affective image annotation. *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, 529–537.