Soft Metrics for Ordinal Label Uncertainty with Variable Sample Sizes

ALESSANDRO MAGNANI, Coupang, USA

We address the challenge of handling uncertainty in ordinal label collections with variable annotation counts (from two to hundreds per instance). Current approaches often ignore both label ordinality and uncertainty from inconsistent sample sizes. We propose two key extensions to existing soft metrics: (1) expectation-based versions of fundamental metrics (KL divergence, cross-entropy, Jensen-Shannon) under a Dirichlet distribution assumption, with analytical solutions for the KL case, and (2) Earth Mover's Distance as a soft metric to naturally capture ordinal relationships. Our approach is validated on both theoretical examples and real-world datasets, revealing that traditional metrics and our uncertainty-aware measures lead to different model rankings. This framework provides a principled way to handle label uncertainty and ordinality across domains with inconsistent annotation counts. Our implementation is available on GitHub. These advances are particularly valuable for software quality assessment and service evaluation contexts, where confident judgments from varying amounts of feedback are crucial for effective service improvement.

CCS Concepts: • **Computing methodologies** → **Uncertainty quantification**; • **Computer systems organization** → *Reliability*; • **Software and its engineering** → *Software verification and validation.*

Additional Key Words and Phrases: ordinal labels, uncertainty quantification, soft metrics, Dirichlet distribution, Earth Mover's Distance, software evaluation, service quality metrics

ACM Reference Format:
Alessandro Magnani. 2025. Soft Metrics for Ordinal Label Uncertainty with Variable Sample Sizes. In *Proceedings of 6th Asia Service Sciences and Software Engineering Conference (ASSE 2025).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The collection of human annotations for machine learning tasks frequently results in datasets with ordinal labels - ratings or assessments that follow a natural order [1]. Examples abound across domains: e-commerce product ratings (1-5 stars), content moderation severity levels (benign to severely harmful), medical diagnosis classifications (from normal to severe), or educational assessment scores. While these labels are often treated as categorical variables for simplicity, this approach discards valuable ordinal information - a 4-star rating is not just different from a 2-star rating, it is better by a quantifiable amount.

While these ordinal label challenges appear across domains, they are particularly critical in service evaluation contexts. Software quality assessment, user experience ratings, and service satisfaction metrics all rely on ordinal scales that directly impact business decisions. In software engineering, ordinal assessments determine feature prioritization, guide quality assurance efforts, and influence release decisions. The methodology we propose addresses a fundamental challenge in service evaluation: how to reliably interpret human judgments when sample sizes vary dramatically, as is common in real-world service feedback systems.

Author's Contact Information: Alessandro Magnani, ale.magnani@gmail.com, Coupang, Mountain View, CA, USA.

Manuscript submitted to ACM

A fundamental challenge in working with such datasets is the distinction between hard and soft labels. Hard labels assign a single categorical value to each instance, while soft labels maintain a probability distribution over possible values [2]. While hard labels are simpler to work with computationally, they can be misleading when there is genuine uncertainty or disagreement among annotators [3].

The challenge is compounded by the practical reality of data collection: the number of annotations per instance often varies dramatically within the same dataset. Some instances might have hundreds of ratings (popular products on e-commerce sites), while others might have as few as two (newly listed items). This variable sample size introduces different levels of uncertainty in our understanding of the true label distribution - we should be more confident in the average rating from 100 reviewers than from 2 reviewers [3].

To address these challenges, we propose a principled framework that makes three key contributions:

(1) We develop expectation-based versions of fundamental soft metrics (KL divergence, cross-entropy, and Jensen-Shannon divergence) under a Dirichlet distribution assumption. This Bayesian approach naturally handles varying sample sizes by appropriately weighting the influence of prior knowledge versus observed data.
(2) We introduce Earth Mover's Distance (EMD) as a soft metric that explicitly captures ordinal relationships between labels. Unlike traditional metrics that treat categories as unrelated, EMD accounts for the distance between ordinal levels, providing more meaningful comparisons for ordinal data.
(3) We validate our approach through comprehensive experiments on both synthetic examples and two real-world datasets: Amazon Electronics Reviews [4] and ConvAbuse [5]. These experiments demonstrate that our uncertainty-aware metrics lead to systematically different model rankings compared to traditional approaches, particularly when annotation counts vary.

Our framework offers several key advantages over existing approaches:

- It naturally handles datasets with varying numbers of annotations per instance, appropriately adjusting confidence based on sample size
- It preserves and leverages ordinal relationships between labels, providing more meaningful metrics for ordinal data
- It provides well-calibrated uncertainty estimates that can be incorporated into downstream training procedures
- It maintains computational efficiency through analytical solutions and optimized approximations
- **Software Quality Assessment:** Our framework provides more reliable evaluation of software features and components when assessments come from varying numbers of testers or users
- **Service Feedback Analysis:** The approach enables more accurate interpretation of customer satisfaction data across services with different usage levels

Traditional approaches to handling label uncertainty have centered around three key soft metrics: Cross-entropy, KL divergence, and Jensen-Shannon divergence [2]. These metrics allow us to compare predicted probability distributions with target distributions, but they typically assume the target distribution is known with certainty. In reality, when we have only a small sample of annotations, there is uncertainty in the target distribution itself [6]. Our work addresses this limitation by explicitly modeling the uncertainty in the target distribution while preserving ordinal relationships between labels.

To facilitate research reproducibility and adoption in practical applications, we provide a comprehensive implementation of all proposed metrics and experimental code at https://github.com/alemagnani/label_uncertainty. This includes utilities for working with both the Amazon Electronics dataset and the ConvAbuse dataset, as well as implementations of the baseline models used in our experiments.

2 Related Work

2.1 Background on Label Uncertainty and Ordinal Relationships

The challenge of human label variation has been increasingly recognized as a fundamental aspect of machine learning systems. [1] frames this as an inherent characteristic rather than a problem to be solved, while [6] found genuine disagreement in at least 20% of natural language inference annotations.

[3] categorize label variation into genuine disagreement from ambiguity, subjective differences in interpretation, and cases allowing multiple plausible answers. Traditional approaches either attempt to resolve uncertainty through aggregation [7, 8] or embrace uncertainty as informative signal [9, 10]. The latter has gained traction in domains with inherent subjectivity such as emotion detection [11] and content moderation.

Statistical approaches to label uncertainty have evolved from early probabilistic methods [12] to more sophisticated Bayesian approaches [13, 14]. Our work builds on [3]'s framework for learning from disagreement, extending it to handle ordinal relationships and variable sample sizes.

The special nature of ordinal labels has been recognized across domains including medical imaging [15] and sentiment analysis [16], with researchers criticizing the common treatment of ordinal labels as categorical variables [17, 18]. Our application of Earth Mover's Distance to label distributions and integration with Bayesian uncertainty modeling represents a novel contribution.

2.2 Approaches to Label Uncertainty

The statistical treatment of label uncertainty has evolved significantly. [12] pioneered early probabilistic approaches, but recent work has introduced more sophisticated methods. [2] introduced human uncertainty modeling using Gaussian processes, while Bayesian approaches have been developed by several researchers [13, 14].

Our work builds particularly on [3]'s framework for learning from disagreement, extending it to handle ordinal relationships and variable sample sizes. The use of Dirichlet distributions for modeling label uncertainty has precedent in the work of [19], though our application to soft metrics is novel.

2.3 Ordinal Labels in Machine Learning

The special nature of ordinal labels has been recognized in various domains. [15] demonstrated their importance in medical image annotation, while [16] explored their role in sentiment analysis. The traditional treatment of ordinal labels as categorical variables has been criticized by several researchers [17, 18].

Our approach to handling ordinal relationships using Earth Mover's Distance builds on work by [20] in the context of image comparison. However, our application to label distributions and integration with Bayesian uncertainty modeling represents a novel contribution.

2.4 Applications in E-commerce and Content Moderation

The practical importance of our work is particularly evident in e-commerce and content moderation settings. [5] demonstrated the challenges of handling nuanced abuse detection annotations, while [21] explored similar issues in product rating systems. These domains frequently encounter both variable numbers of annotations and inherently ordinal labels, making them ideal applications for our methods.

2.5 Approaches to Label Uncertainty

The challenge of learning from uncertain or ambiguous labels has been approached from multiple perspectives. Peterson et al. [2] proposed Joint-Distribution Soft Loss (JDSL), which directly models human uncertainty through soft label distributions rather than forcing consensus to hard labels. Their work shows that preserving disagreement information during training results in models that are more robust to annotation noise, though they don't specifically address varying sample sizes or ordinal relationships.

Learning from Label Proportions (LLP) [22] presents a related approach where only the proportions of labels within groups are available. While originally developed for weakly supervised settings, these techniques can be adapted to scenarios where each item has multiple annotations. However, LLP methods typically don't account for the uncertainties introduced by small sample sizes.

The crowdsourcing literature offers several approaches to annotation uncertainty. The classic Dawid-Skene model [7] and its extensions by Raykar et al. [23] handle multiple annotators with varying reliability, but these methods focus primarily on estimating annotator quality rather than modeling the uncertainty from limited sample sizes. Paun et al. [24] provide a comprehensive comparison of aggregation methods for annotations, though they primarily target categorical rather than ordinal labels.

Bayesian approaches to label uncertainty have gained prominence recently. Kendall and Gal [25] distinguish between aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model uncertainty) in deep learning, with the former being particularly relevant to our problem of label noise. However, their approach doesn't specifically address the ordinal nature of labels.

Cui et al. [26] introduced Distribution Matching Networks that use Wasserstein distances for learning from frequency distributions, making them relevant to our use of Earth Mover's Distance for ordinal labels. Their work, however, doesn't incorporate Bayesian uncertainty quantification based on sample size.

Fornaciari et al. [10] demonstrated the value of learning directly from disagreement distributions rather than aggregated labels, but did not address the variable confidence resulting from different sample sizes. Similarly, Uma et al. [3] proposed a framework for learning from disagreement but without the Bayesian calibration we propose.

For handling specifically ordinal labels, traditional approaches include ordered probit/logit models [27], while more recent machine learning approaches include Gaussian process ordinal regression [28] and ordinal distribution regression [29]. These methods capture the ordinal relationships but typically don't incorporate uncertainty from variable annotation counts.

Our work differs from these approaches by simultaneously addressing both the ordinal nature of labels and the uncertainty arising from variable sample sizes within a unified Bayesian framework. While some previous approaches have addressed one aspect or the other, to our knowledge, no existing method provides the integrated treatment we propose.

3 Methodology

3.1 Problem Formulation

Consider a rating system with K ordered labels $\{0, ..., K-1\}$ where each label has a semantic interpretation (e.g., for $K = 5$: 0="poor", 1="fair", 2="good", 3="very good", 4="excellent"). This corresponds to a 1-5 star rating system commonly used in applications. For each item i, we collect n_i ratings, where n_i may vary across items.

3.2 Bayesian Label Distribution Modeling

Given a collection of labels with varying levels of annotator agreement and different sample sizes per instance, we need to model the inherent uncertainties in these labels. We approach this by formulating a Bayesian framework that explicitly models two key sources of uncertainty: the uncertainty in individual label assignments and the uncertainty in the underlying probability distribution of labels. This principled approach allows us to both incorporate prior knowledge and properly account for varying numbers of annotations per instance, while maintaining a clear understanding of our confidence in each label distribution.

3.2.1 Prior and Posterior Distributions. For a K-class labeling problem, we model the prior distribution over label probabilities $p = (p_1, ..., p_K)$ using a Dirichlet distribution:

$$p \sim \mathrm{Dir}(\alpha_1, ..., \alpha_K) \tag{1}$$

where $(\alpha_1, ..., \alpha_K)$ are the prior concentration parameters. A natural uninformative prior is the uniform Dirichlet with $\alpha_i = 1$ for all i, though the framework allows for informative priors when there is reason to expect certain labels to be more likely.

Given observed label counts $n = (n_1, ..., n_K)$, the posterior distribution is also Dirichlet due to conjugacy:

$$p|n \sim \mathrm{Dir}(\alpha_1 + n_1, ..., \alpha_K + n_K) \tag{2}$$

3.2.2 Expected Label Probabilities. The expected probability of label i under this model is:

$$E[p_i|n] = \frac{\alpha_i + n_i}{\alpha_0 + n_0} \tag{3}$$

where $\alpha_0 = \sum_i \alpha_i$ and $n_0 = \sum_i n_i$. This formula reveals an important correction to raw frequency estimates. For example, with a uniform prior ($\alpha_i = 1$) and two observed "excellent" ratings out of two total ratings, the expected probability of "excellent" would be:

$$E[p_{\text{excellent}}] = \frac{1+2}{K+2} \tag{4}$$

This is notably different from the raw frequency of 1.0, reflecting our uncertainty given the small sample size. The correction becomes less significant with larger sample sizes, as the posterior becomes dominated by the observed data.

3.2.3 Expected Metrics. Given this Bayesian formulation, we can derive expected versions of standard soft metrics. For a predicted distribution q, the expected cross-entropy is:

$$E[\mathrm{CrossEntropy}(p, q)] = -\sum_i E[p_i|n] \log(q_i) = -\sum_i \frac{\alpha_i + n_i}{\alpha_0 + n_0} \log(q_i) \tag{5}$$

The full derivation of this formula involves integrating over the Dirichlet distribution and applying properties of the digamma function. For space considerations, we provide the result here while a complete proof will be presented in Section 5.

For comparing distributions, both KL divergence and cross-entropy lead to equivalent model rankings since they differ only by the entropy term $H(p)$, which is independent of the predicted distribution q. The expected KL divergence is:

$$E[\mathrm{KL}(p||q)] = \sum_i \frac{\alpha_i + n_i}{\alpha_0 + n_0} (\psi(\alpha_i + n_i + 1) - \psi(\alpha_0 + n_0 + 1) - \log(q_i)) \tag{6}$$

where $\psi(x)$ is the digamma function.

3.2.4 Expected Metrics and Their Properties. The derived expected metrics offer several advantages: (1) **computational efficiency** despite their Bayesian foundation, requiring only replacement of raw frequencies with expected counterparts; (2) **analytical tractability** with clean closed-form solutions; (3) **smooth uncertainty transition** between prior knowledge and empirical observations; (4) significant **impact on decision boundaries** in practice, especially with few annotations; and (5) well-calibrated **uncertainty estimates** reflecting both annotation randomness and distributional uncertainty. These properties make our framework suitable for applications with varying annotation counts, combining theoretical elegance with practical efficiency.

3.3 Earth Mover's Distance for Ordinal Labels

The Earth Mover's Distance (EMD), also known as the Wasserstein metric, provides a natural way to capture the ordinal relationship between labels. Unlike traditional metrics that treat labels as categorical variables, EMD incorporates the "distance" that must be traveled to transform one distribution into another.

In our ordinal rating context with K labels $\{0, ..., K-1\}$, we define the ground distance between labels i and j as:

$$d(i,j) = \frac{|i-j|}{K-1} \tag{7}$$

This normalization ensures distances are in $[0,1]$ while preserving the ordinal relationships. The EMD between two distributions P and Q is defined as:

$$\mathrm{EMD}(P,Q) = \min_F \sum_{i,j} f_{ij} d(i,j) \quad \text{s.t.} \quad f_{ij} \geq 0, \quad \sum_j f_{ij} = P(i), \quad \sum_i f_{ij} = Q(j) \tag{8}$$

where f_{ij} represents the flow from bin i to bin j, subject to non-negative flow and mass preservation constraints.

3.3.1 Expected EMD under Dirichlet Prior. While we analytically derive expectations for KL divergence under a Dirichlet prior, the EMD expectation $E[\mathrm{EMD}(p,q)] = \int \mathrm{EMD}(p,q)\mathrm{Dir}(p|\alpha)dp$ lacks a closed-form solution. We approximate it through Monte Carlo integration: $E[\mathrm{EMD}(p,q)] \approx \frac{1}{N}\sum_{i=1}^{N} \mathrm{EMD}(p^{(i)},q)$, where $p^{(i)}$ are Dirichlet samples. Moderate values of N (e.g., 1000) provide stable estimates.

3.3.2 Computational Complexity. For standard EMD: (1) Basic computation: $O(K^3 \log K)$, reduced to $O(K)$ for 1D ordered labels through cumulative distribution differences, with $O(K^2)$ memory for the cost matrix. For expected EMD: Monte Carlo with N samples requires $O(NK)$ time. Practical optimizations include pre-computing cost matrices, parallel computation, and caching samples for repeated computations.

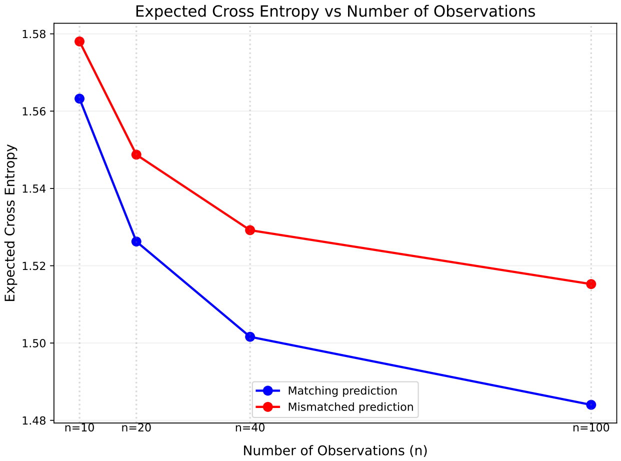

Fig. 1. Expected cross-entropy for matching and mismatched predictions as observations increase, showing reduced uncertainty and increased discrimination with more data.

3.4 Illustrative Examples

3.4.1 Example 1: Impact of Sample Size. Consider a 5-star rating system with: initial observations [1,2,4,2,1] (centered at 2 stars), uniform Dirichlet prior $\alpha = [1, 1, 1, 1, 1]$, and two predictive distributions: Matching $q = [0.1, 0.2, 0.4, 0.2, 0.1]$ and Mismatched $q = [0.1, 0.3, 0.3, 0.2, 0.1]$. Figure 1 shows how expected cross-entropy changes as we increase observations by factors of 1× (10 ratings), 2× (20), 4× (40), and 10× (100). The expected cross-entropy is:

$$E[\text{CrossEntropy}(p, q)] = -\sum_{i=1}^{K} E[p_i] \log(q_i) = -\sum_{i=1}^{K} \frac{\alpha_i + n_i}{\alpha_0 + n_0} \log(q_i) \tag{9}$$

The plot reveals three key insights: (1) With few observations, cross-entropy is higher for both predictions due to uncertainty, with less pronounced differences between good and poor predictions; (2) As observations increase, cross-entropy decreases as prior uncertainty diminishes; (3) The gap between matching and mismatched predictions widens with more data, showing increased metric discrimination. This example illustrates how our framework naturally handles varying sample sizes, providing appropriate uncertainty estimates that make the metric more conservative with few ratings and more discriminative with many.

3.4.2 Example 2: Ordinal vs Categorical Differences. Consider two scenarios with identical KL divergence but different ordinal implications:

4 Experimental Validation on Amazon Electronics Reviews Dataset

To validate our theoretical framework in a real-world setting, we conducted experiments using the Amazon Electronics Reviews dataset [4]. This experiment provides empirical evidence for our two key contributions: properly accounting for sample size uncertainty and preserving ordinal relationships between labels.

Scenario	A: Adjacent Label Confusion	B: Distant Label Confusion
True distribution	[0.0, 0.8, 0.2, 0.0, 0.0]	[0.0, 0.8, 0.0, 0.0, 0.2]
Predicted distribution	[0.0, 0.2, 0.8, 0.0, 0.0]	[0.0, 0.2, 0.0, 0.0, 0.8]
Error type	Adjacent categories	Distant categories
Metric Comparison		
KL Divergence	$\mathrm{KL}(A) \approx \mathrm{KL}(B) \approx 0.97$	
Cross-Entropy	$\mathrm{CrossEntropy}(A) \approx \mathrm{CrossEntropy}(B) \approx 1.83$	
EMD (Our metric)	$\mathrm{EMD}(A) \approx 0.15$ (small error)	$\mathrm{EMD}(B) \approx 0.45$ (large error)

Table 1. Comparison of metrics for ordinal label confusion. Traditional metrics (KL and CE) assign equal scores to both scenarios, while EMD correctly identifies the greater ordinal error in scenario B.

Table 2. Overall model ranking by metric type (lower scores are better, best scores in **bold**)

Metric	Model A	Model B	Model C	Model D
CE Empirical	**1.251**	1.255	1.268	1.265
E[CE]	1.455	**1.443**	1.451	1.447
EMD Empirical	**0.531**	0.536	0.544	0.534
E[EMD]	0.542	**0.526**	0.542	0.532

4.1 Experimental Setup

We sampled 20,000 products with at least 5 reviews each, using product titles as features for predicting rating distributions. Products exhibited natural variation in review counts, ranging from as few as 5 reviews to over 100 per product. We implemented and compared four distinct models:

- **Model A (MLP+CE)**: A standard multilayer perceptron trained with categorical cross-entropy loss against raw empirical distributions.
- **Model B (MLP+KL)**: Identical architecture to Model A, but trained with KL divergence loss on Gaussian-smoothed target distributions, implicitly respecting ordinal relationships.
- **Model C (BONN)**: A Bayesian Ordinal Neural Network explicitly modeling the ordinal nature of ratings through a cumulative link approach.
- **Model D (TORP)**: A transformer-based architecture trained with EMD-aware loss, representing a more complex approach.

4.2 Results and Analysis

Our experimental results reveal three key findings that support our theoretical framework:

4.2.1 Metric Choice Determines Model Selection. The most striking result, shown in Table 2, is how the choice of evaluation metric dramatically affects which model appears superior. Under traditional empirical metrics (CE_emp and EMD_emp), the simple MLP with standard cross-entropy loss (Model A) outperforms all others. However, when evaluated with expected metrics that incorporate uncertainty, Model B emerges as the clear winner.

This model ranking reversal confirms our central claim: failing to account for label uncertainty leads to systematically different evaluation outcomes. The smoothed-target approach used in Model B implicitly respects both uncertainty and ordinal relationships, explaining its superior performance under metrics that value these properties.

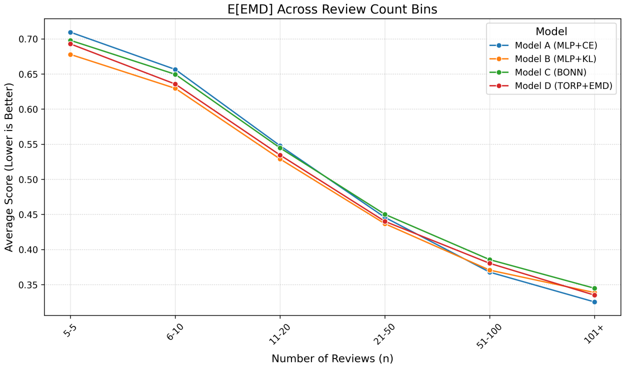

Fig. 2. E[EMD] across review count bins shows Model B outperforms others with fewer reviews.

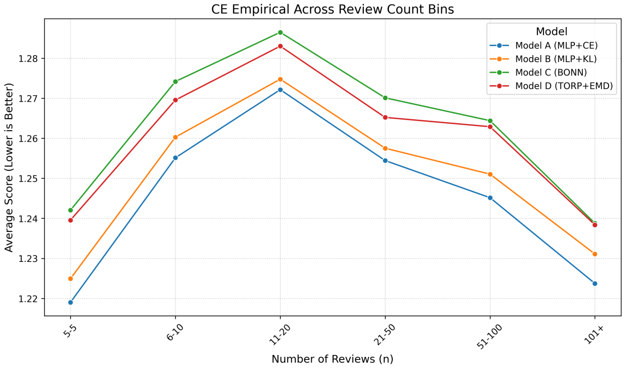

Fig. 3. CE Empirical scores show non-monotonic relationship with review counts, highlighting limitations of empirical metrics with variable sample sizes.

4.2.2 Uncertainty Calibration by Sample Size. Figure 2 demonstrates how model performance varies with sample size when evaluated using E[EMD]. For products with minimal reviews (5-5 bin), Model B shows a clear advantage over other approaches. As review counts increase, this advantage diminishes, and by the 101+ bin, the gap becomes minimal.

This pattern directly confirms our theoretical prediction: uncertainty modeling matters most when sample sizes are small. The fact that Model B's advantage is most pronounced in the 5-5 bin validates our framework's emphasis on properly calibrating confidence based on available evidence.

4.2.3 The Paradox of Empirical Metrics. An unexpected finding shown in Figure 3 is the non-monotonic relationship between review count and empirical CE performance. For all models, performance initially worsens as review counts increase from 5 to 20, then improves again for products with more reviews. This counterintuitive pattern highlights a fundamental challenge with empirical metrics: they don't account for the higher variance in small samples.

When a product has only 5 reviews, empirical metrics "reward" models for matching the observed distribution exactly, even though this distribution is likely a poor estimate of the true underlying preferences. Our expected metrics avoid this pitfall by appropriately discounting noisy empirical distributions from small samples.

The fact that sophisticated architectures (Models C and D) don't outperform the simpler Model B when evaluated with uncertainty-aware metrics indicates that the training objective—specifically, using KL divergence with smoothed targets—may be more important than architectural complexity when handling ordinal label uncertainty.

4.3 Practical Implications

These results have significant practical implications for e-commerce platforms and other applications involving ordinal ratings:

1. **Model Selection Impact**: Using expected metrics leads to selecting different models than traditional empirical metrics would suggest, potentially improving performance on new items with few ratings.
2. **Training Strategy**: Smoothing target distributions during training (as in Model B) effectively captures ordinal relationships without requiring complex architectures.
3. **Sample Size Adaptation**: Expected metrics naturally adjust confidence based on available evidence, providing more reliable evaluations across different sample size regimes.

The model ranking reversal between empirical and expected metrics confirms that traditional evaluation approaches may systematically undervalue models that handle uncertainty appropriately. For real-world recommendation systems dealing with variable review counts, our framework provides a principled approach to both training and evaluation that respects the confidence we should place in distributions from different sample sizes.

These experimental results on real-world data validate both key contributions of our work: properly accounting for varying sample sizes and preserving ordinal relationships between labels.

5 Experimental Validation on ConvAbuse Dataset

To further validate our framework in a different domain, we conducted experiments using the ConvAbuse dataset [5], which contains annotated conversations between users and three conversational AI systems. This experiment provides empirical evidence for our approach's effectiveness in handling both label uncertainty and ordinal relationships in a conversational abuse detection context.

5.1 Dataset Characteristics

The ConvAbuse dataset comprises conversations from three different conversational systems: an open-domain social bot (Alana v2), a rule-based chatbot (ELIZA), and a task-based system (CarbonBot). The dataset includes fine-grained annotations of abuse severity on an ordinal scale from +1 (non-abusive) to -3 (strongly abusive), provided by multiple expert annotators with backgrounds in gender studies. Unlike the Amazon Electronics dataset where uncertainty stems from naturally varying numbers of product ratings, the ConvAbuse dataset features a more controlled annotation environment with each instance typically having 2-5 expert annotations.

The dataset's ordinal nature and expert annotations provide an ideal test case for our metrics that are designed to handle ordinal relationships while accounting for annotation uncertainty.

5.2 Experimental Setup

For this experiment, we implemented two BERT-based models:

- **Model A (No Context)**: A model trained on individual user utterances without considering the surrounding conversational context.

Table 3. Performance on ConvAbuse dataset (lower is better, best in **bold**)

Metric	A	B	Winner	Metric	A	B	Winner
CE Emp	0.473	0.484	A	EMD Emp	**0.251**	0.276	A
E[CE]	2.580	**2.526**	B	E[EMD]	1.125	**1.108**	B

Table 4. Average scores per annotation count bin (lower is better)

n_bin	Count	CE_emp		E[CE]	
		A	B	A	B
2-3	508	**0.498**	0.513	2.649	**2.588**
4-5	93	0.358	**0.351**	2.248	**2.229**
6-10	8	**0.199**	0.202	2.055	**2.047**

- **Model B (With Context)**: An identical architecture trained with conversational context, including previous turns from both the user and the system.

We applied our Dirichlet-based uncertainty handling approach to appropriately weight the annotations, setting uniform priors with $\alpha = [1, 1, 1, 1, 1]$ as in our previous experiments. We evaluated both models using traditional empirical metrics (CE and EMD calculated on raw annotation frequencies) and our expected metrics that incorporate uncertainty (E[CE] and E[EMD]).

5.3 Results and Analysis

Our experimental results, summarized in Table 3, reveal a striking pattern that aligns with our findings from the Amazon Electronics dataset. When evaluated using traditional empirical metrics, Model A (without context) appears to outperform Model B. However, when our expected metrics that incorporate uncertainty are applied, the ranking is reversed, with Model B (with context) emerging as the superior model.

This reversal in model ranking provides further evidence for our central thesis that traditional metrics may systematically undervalue models that handle uncertainty appropriately. The contextual information utilized by Model B provides valuable signals for abuse detection, but this benefit is only apparent when uncertainty is properly modeled using our expected metrics.

To better understand the effect of annotation count on model performance, we analyzed results across different bins of annotation counts, as shown in Table 4. The analysis reveals several key patterns:

1. **Decreasing metric values with increasing annotation counts**: Across all metrics, scores decrease (improve) as the number of annotations increases, reflecting lower uncertainty with more annotator consensus.
2. **Model B's advantage in expected metrics**: In instances with few annotations (2-3), Model B shows a clear advantage under the expected metrics, despite being outperformed by Model A under empirical metrics for the same instances.
3. **Converging performance with more annotations**: As annotation counts increase, the performance gap between the two models narrows, suggesting that with sufficient annotations, even empirical metrics begin to capture the benefits of contextual information.

The detailed bin-by-bin analysis shows that the majority of the test data (508 out of 609 instances) falls into the 2-3 annotation bin, highlighting the prevalence of scenarios with minimal annotations in real-world applications. This underscores the importance of methods that can handle uncertainty from limited annotations.

5.4 Discussion

These results on the ConvAbuse dataset complement our findings from the Amazon Electronics experiments in several important ways:

- **Domain Generalization**: The consistent pattern of model ranking reversal between empirical and expected metrics across two vastly different domains—e-commerce product ratings and conversational abuse detection—demonstrates the broad applicability of our approach.
- **Annotation Paradigm Transfer**: While the Amazon dataset featured naturally occurring ratings with highly variable counts (from a few to hundreds), ConvAbuse represents a controlled annotation environment with expert judgments (typically 2-5 per instance). Our framework proves effective in both paradigms.
- **Value of Contextual Information**: The ConvAbuse results suggest that conversational context provides substantial information for abuse detection, but this benefit is only apparent when uncertainty is properly modeled using our expected metrics. This demonstrates how inappropriate uncertainty handling can mask genuine model improvements.
- **Practical Implications**: In abuse detection scenarios where obtaining large numbers of annotations is costly or impractical, our expected metrics provide a more reliable evaluation framework that correctly identifies superior models even with minimal annotations.

The Earth Mover's Distance metrics successfully capture the ordinal nature of abuse severity ratings, with E[EMD] providing a clear differentiation between models that aligns with our expected result that including context should improve abuse detection. This further validates our approach to handling uncertain ordinal labels, showing its effectiveness in the important application area of conversational abuse detection.

6 Derivation of Expected Entropy

Under the Dirichlet distribution, the expectation becomes:

$$E[p_i \log p_i] = \int p_i \log p_i \frac{1}{B(\alpha)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} dp \tag{10}$$

where $B(\alpha)$ is the multivariate beta function.

This integral can be solved by recognizing that it's related to the derivative of the log normalization constant with respect to α_i. The key insight is:

$$E[p_i \log p_i] = \frac{\partial}{\partial \alpha_i} \log B(\alpha) \tag{11}$$

The derivative of the log multivariate beta function can be expressed in terms of the digamma function:

$$\frac{\partial}{\partial \alpha_i} \log B(\alpha) = \frac{\alpha_i}{\alpha_0} (\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)) \tag{12}$$

where $\psi(x)$ is the digamma function and $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. Therefore, the expected entropy is:

$$E[H(p)] = -\sum_{i=1}^{K} \frac{\alpha_i}{\alpha_0} (\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)) \tag{13}$$

6.1 Properties of the Result

This result exhibits key properties: it accounts for both individual α_i values and their sum; converges to expected distribution entropy as $\alpha_0 \to \infty$ with fixed ratios; reaches maximum uncertainty with uniform prior ($\alpha_i = 1$); maintains non-negativity; and preserves symmetry in parameters, respecting Dirichlet exchangeability.

7 Implementation Considerations for Service Systems

The metrics and methods described in this paper can be readily implemented in existing service evaluation systems. Our framework requires minimal modifications to standard evaluation pipelines and can be deployed alongside traditional metrics for comparison. For software development teams, we recommend implementing these metrics in quality assurance dashboards to provide more nuanced evaluation of features with varying amounts of feedback. This can be particularly valuable during beta testing phases when feedback volume is inconsistent.

For service scientists, our GitHub implementation includes tools for analyzing service ratings across multiple channels with varying feedback volumes. The codebase is modular and integrates with standard data analysis workflows, requiring minimal adaptation for production deployment. Most implementations require only a few dozen lines of additional code to incorporate our Bayesian uncertainty handling approach.

8 Conclusions

Our framework addresses the challenge of handling uncertainty in ordinal label collections with varying sample sizes, providing three key contributions: expectation-based soft metrics, ordinal-aware distance measures, and analytical solutions for uncertainty quantification. These advances directly benefit applications including e-commerce rating predictions, content moderation, and medical diagnostics.

For service sciences and software engineering practitioners, this work provides immediate practical value. Software teams can apply our metrics to more accurately interpret user feedback during beta testing, where early features may have few ratings while established features have many. The computational efficiency allows for seamless integration into existing evaluation pipelines.

Our approach eliminates the false dichotomy between hard and soft labels while preserving ordinal relationships. Experiments on synthetic examples and real-world datasets consistently demonstrate that traditional metrics and our uncertainty-aware measures lead to different model rankings, with significant implications for model selection.

The implementation on GitHub facilitates research reproducibility and practical adoption. Future work could extend this framework to hierarchical label structures, domain-specific prior knowledge, and integration with active learning systems. Our work highlights the value of treating human label variation as a signal to be understood rather than noise to be eliminated, contributing to more reliable machine learning systems where human judgment plays a crucial role.

References

[1] B. Plank, “The "problem" of human label variation: On ground truth in data, modeling and evaluation,” *arXiv preprint arXiv:2211.02570*, 2022.

[2] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, “Human uncertainty makes classification more robust,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9617–9626.

[3] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, “Learning from disagreement: A survey,” *Journal of Artificial Intelligence Research (JAIR)*, vol. 72, pp. 1385–1470, 2021.

[4] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” in *Proceedings of the 25th International Conference on World Wide Web (WWW)*. ACM, 2016, pp. 507–517.

[5] A. Cercas Curry, G. Abercrombie, and V. Rieser, “ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online and Punta Cana, Dominican Republic:

Association for Computational Linguistics, Nov. 2021, pp. 7388–7403. [Online]. Available: https://aclanthology.org/2021.emnlp-main.587

[6] E. Pavlick and T. Kwiatkowski, "Inherent disagreements in human textual inferences," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 7, pp. 677–694, 2019.

[7] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[8] S. Paun, R. Artstein, and M. Poesio, "Statistical methods for annotation analysis," *Synthesis Lectures on Human Language Technologies*, vol. 15, no. 1, pp. 1–217, 2022.

[9] P. Sommerauer, A. Fokkens, and P. Vossen, "Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020, pp. 4798–4809.

[10] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, and M. Poesio, "Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021, pp. 2591–2597.

[11] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4040–4054.

[12] P. Smyth, "Inference in probabilistic expert systems," *Statistical Science*, vol. 10, no. 1, pp. 48–58, 1995.

[13] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*, 2021, pp. 1–14.

[14] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1910–1918.

[15] V. Cheplygina and J. P. Pluim, "Crowd disagreement about medical images is informative," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI Workshop)*, ser. Lecture Notes in Computer Science, vol. 11037. Springer, 2018, pp. 105–111.

[16] C. O. Alm, "Subjective natural language problems: Motivations, applications, characterizations, and implications," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011, pp. 107–112.

[17] C. D. Manning, "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" in *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, ser. Lecture Notes in Computer Science, vol. 6608. Springer, 2011, pp. 171–189.

[18] D. Zeman, "Hard problems of tagset conversion," in *Proceedings of the Second International Conference on Global Interoperability for Language Resources (GIGR)*, 2010, pp. 181–185.

[19] C. Archambeau, "An introduction to probabilistic models with missing data," in *Statistical Data Analysis for the Physical Sciences*, G. D'Agostini, Ed.

[20] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2001, pp. 251–256.

[21] T. Liu, C. Homan, C. O. Alm, A. M. White, M. C. Lytle, and H. A. Kautz, "Understanding discourse on work and job-related well-being in public social media," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 1044–1053.

[22] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 847–854.

[23] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 1297–1322, 2010.

[24] S. Paun, B. Carpenter, J. Chamberlain, D. Hovy, U. Kruschwitz, and M. Poesio, "Comparing Bayesian models of annotation," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 6, pp. 571–585, 2018.

[25] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.

[26] S. Cui, X. Chen, J. Wen, H. Sun, D. Lei, and L. Shi, "Learning from label proportions with consistency regularization," in *Asian Conference on Machine Learning (ACML)*, ser. Proceedings of Machine Learning Research, vol. 129. PMLR, 2020, pp. 500–515.

[27] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[28] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 1019–1041, 2005.

[29] M. Seeger, Y.-X. Zhang, L. Van Der Maaten, and H. Wang, "Ordinal distribution regression for gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 780–12 789.