# Capturing Epistemic Uncertainty in LLM-Based Soft Labeling

**Yanru Jiang***
Department of Communication
Department of Statistics
University of California, Los Angeles
yanrujiang@ucla.edu

**Siyu Liang***
Department of Political Science
Department of Statistics
University of California, Los Angeles
sliang46@ucla.edu

## Abstract

In many human-annotated NLP tasks involving ambiguity or subjective judgment, annotator disagreement reflects epistemic uncertainty rather than noise. Soft labeling (SL), which represents annotations as probability distributions rather than majority-vote (MV) labels, preserves this uncertainty and can improve downstream performance. We extend this perspective to LLM-based annotation by formalizing LLM soft labeling as introducing controlled variation in model-generated annotations to approximate the latent variability underlying human annotations. We distinguish two sources of variation: *model-induced* (e.g., stochastic decoding and model ensembles) and *human-approximated* (e.g., persona prompting and human-calibrated in-context annotation). Using the Gab Hate and GoEmotions datasets, we show that SL training consistently outperforms MV training under stronger LLM-based annotation strategies. Model ensembles produce the most informative soft-label distributions, achieving the best human–LLM agreement and downstream classification performance. These findings suggest that scalable LLM-based annotation pipelines can model epistemic uncertainty through diverse model-level variation without explicitly simulating human attributes.

## 1 Introduction

Annotated data underpin many machine learning tasks, including classification, ranking, retrieval, and evaluation. Most pipelines assume a single ground-truth label, resolving disagreement via majority vote and collapsing meaningful variation in human judgments into a single "hard" label (Basile et al., 2021).

Recent work in NLP has increasingly questioned this assumption (Uma et al., 2021). In many tasks, such as stance detection, hate speech, or sentiment analysis, texts can support multiple plausible interpretations, and disagreement may reflect genuine ambiguity rather than error. More broadly, mapping high-dimensional linguistic signals into a small set of discrete categories often makes multiple labels defensible for a single instance.

Soft labeling captures this uncertainty by modeling the distribution of annotator judgments, and has been shown to improve performance in ambiguous settings while preserving epistemic uncertainty (Liang, 2026; Li et al., 2025; Meissner et al., 2021). More recently, its applicability extends beyond classification to tasks such as personality modeling (Li et al., 2025) and evaluation of open-ended LLM responses (Jiang et al., 2025), where pluralistic disagreement is preferred and expected.

LLMs are increasingly used as automated annotators, enabling scalable evaluation and labeling (Gu et al., 2024). However, most LLM annotation pipelines still produce a single label, either from one generation or via majority vote across outputs, thereby discarding uncertainty (Davani et al., 2022). Simply sampling multiple outputs may be insufficient, as recent work shows substantial homogeneity within model families (Jiang et al., 2025).

To address this, we formalize LLM-based soft labeling as introducing controlled variation in model-generated annotations to approximate the latent variability underlying human annotations. We distinguish two sources of variation: *model-induced* (e.g., stochastic decoding, model ensembles) and *human-approximated* (e.g., persona prompting, human-calibrated in-context annotation), yielding four LLM-based annotation strategies.

We further introduce a unified evaluation framework spanning soft and hard labeling, combining individual-level agreement and downstream performance metrics applicable to both soft-label and majority-vote training. Using the Gab Hate and GoEmotions datasets, we show that soft-label training consistently outperforms majority-vote under

*Equal contribution.



*Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 177–190

stronger strategies, particularly model ensembles and human-calibrated in-context annotation.

Our main contributions are: (1) we formalize LLM-based soft labeling as controlled variation that approximates latent human annotation variability; (2) we propose a unified evaluation framework compatible with both soft-label and majority-vote training; and (3) we identify model ensembles as the most effective strategy, challenging the assumption that annotation diversity requires explicit simulation of human attributes.

## 2 Related Work

### 2.1 Soft Labeling as Modeling Epistemic Uncertainty

Soft labeling has been applied in a range of domains, including emotion detection (Washington et al., 2021), hate speech detection (Davani et al., 2022), COVID-related misinformation and stance detection (Wu et al., 2023), image classification (Collins et al., 2022; Peterson et al., 2019), and clinical data analysis (de Vries and Thierens, 2025). Across these studies, models trained on soft labels often exhibit better calibration and robustness, and in some cases improved predictive performance (e.g., higher F1 and lower cross-entropy), compared with models trained on majority-vote labels (Collins et al., 2022; de Vries and Thierens, 2025; Davani et al., 2022; Yuan et al., 2023).

A growing body of work develops methods to learn from soft labels. For example, Fornaciari et al. (2021) propose a multi-task learning framework that predicts label distributions alongside standard classification outputs. Davani et al. (2022) model each annotator as a separate task while sharing a common representation of the input text, enabling models to directly learn from annotator disagreement. Wang and Yoon (2022) incorporate soft labels through knowledge distillation, where probabilistic outputs from a teacher model provide richer supervision for training. Collins et al. (2022) collect probabilistic judgments from annotators to construct confidence-based soft labels across multiple classes. Wu et al. (2023) propose a Bayesian confidence calibration framework that improves classification performance by incorporating annotator confidence and secondary labels. Overall, these studies highlight the potential of soft labeling as a principled approach for capturing annotation uncertainty and improving both model calibration and robustness (Meissner et al., 2021).

### 2.2 LLM-Based Annotation

Recent advances in LLMs have enabled their use as scalable substitutes for human annotators. Compared with human annotation, LLM labeling can be substantially more efficient while achieving performance comparable to that of human coders (Gilardi et al., 2023; Tan et al., 2024; Wang et al., 2024). Prior research shows that LLMs perform well across a range of tasks, including stance detection (Gilardi et al., 2023), sentiment analysis (Wu et al., 2024), event extraction (Chen et al., 2024), and topic classification (Alizadeh et al., 2025; Tekumalla and Banda, 2023). These applications span multiple domains, including computer vision (Sapkota et al., 2025; Yamagata and Yamada, 2024), public health (Goel et al., 2023; Tekumalla and Banda, 2023), and social sciences (Gilardi et al., 2023; Horych et al., 2025). Additionally, studies also demonstrate that LLMs can produce reliable annotations under both zero-shot (Tekumalla and Banda, 2023; Sapkota et al., 2025) and few-shot prompting (Ahmed et al., 2025; Espinosa and Salathé, 2024; Goel et al., 2023; Wu et al., 2024) settings, highlighting their effectiveness even when limited task-specific training data are available. Recent work further shows that persona-prompted LLMs can increase annotation diversity, and that persona effects on LLM annotations align with subjective variation observed in human judgments (Fröhling et al., 2025).

In most studies, however, LLM-generated annotations are collapsed into single labels through majority vote before downstream use (Choi et al., 2024; Horych et al., 2025; Vallejo Vera and Driggers, 2025; Wang et al., 2024), discarding variation that may reflect uncertainty or multiple plausible interpretations. While some approaches explore diversity through persona prompting (Fröhling et al., 2025; Lee et al., 2026), a systematic framework for modeling and evaluating such variation in LLM-based annotation remains underdeveloped.

## 3 LLM-Based Soft Labeling

Human soft labeling arises from variation across annotators. In traditional annotation pipelines, each document is labeled by multiple individuals, and the resulting distribution of judgments captures epistemic uncertainty across annotators.

Formally, given annotations from a set of annotators $a \in A$, the empirical soft-label distribution can be written as:

$$p_{\text{human}}(y \mid x) = \frac{1}{|A|} \sum_{a \in A} \mathbf{1}[y_a = y],$$

where $y_a$ denotes the label assigned by annotator $a$. This empirical distribution can also be interpreted as arising from latent differences across annotators (Davani et al., 2022). Let $h$ denote latent annotator characteristics (e.g., experience, bias, or demographics). The human annotation distribution can then be written as:

$$p_{\text{human}}(y \mid x) = E_{h \sim p(h)} \left[ \mathbf{1}[g(x, h) = y] \right],$$

where $g(x, h)$ denotes the label assigned by an annotator with characteristics $h$, and $p(h)$ represents the distribution of perspectives within the annotator population.

Our LLM formulation mirrors this perspective by replacing latent human annotator characteristics $h$ with a generic variation variable $z$. Similarly, the LLM soft-label distribution can be defined as:

$$p_{\text{LLM}}(y \mid x) = E_{z \sim q(z)} \left[ \mathbf{1}[f_\theta(x, z) = y] \right],$$

where $f_\theta(x, z)$ denotes the label predicted by the model for document $x$ under variation condition $z$, and $q(z)$ represents the distribution over variation sources. While some LLM-based annotation studies attempt to design $z$ to correspond to human factors (e.g., personality or demographic personas) (Fröhling et al., 2025), our formulation and empirical results suggest that $z$ does not need to directly represent human attributes.

Previous work suggests that labeling disagreement can arise from multiple sources of variation in human judgments (Plank, 2022). Similarly, $z$ can represent any mechanism that induces variation in model labeling behavior, such as stochastic decoding, model ensembles, or prompt variations.

Based on this framework, we examine four LLM-based soft labeling strategies. Our methods introduce variation through two sources. The first source is *model-induced variation*, which arises from stochastic decoding or differences across model families. The second source is *human-approximated variation*, which attempts to approximate systematic differences observed in human annotators, such as demographic perspectives or individual labeling tendencies. We introduce the four approaches here, with implementation details provided in Section 5.3.

### 3.1 Model-Induced Variation

The first class of approaches introduces variation through properties of the model itself.

**Single-Model Repeated Sampling.** Multiple annotations are generated by repeatedly sampling from the same LLM using stochastic decoding, with variation reflecting the model's output distribution. High-temperature sampling is a common baseline in LLM-based annotation diversity studies (Wang et al., 2025).

**Model Ensembles.** Annotations are generated from multiple LLMs and aggregated to simulate inter-annotator disagreement arising from differences across models. Ensembles have been utilized as a promising approach for predicting different annotators' labels (Davani et al., 2022).

### 3.2 Human-Approximated Variation

The second class of approaches aims to approximate structured sources of variation observed in human annotation.

**Demographic Persona Prompting.** Models are prompted with demographic personas that condition responses on group-level attributes such as ideology, age, gender, or education, which prior studies have considered relevant factors in annotation judgments (Spinde et al., 2021). Previous studies have explored various persona prompting strategies, including personality traits (Lee et al., 2026; Jiang and Akçakır, 2025), demographic attributes (Fröhling et al., 2025), as well as multilingual prompting with culturally associated cues, to enhance diversity in annotations and social norm judgments (Wang et al., 2025).

**Human-Calibrated In-Context Annotation.** Models are conditioned on examples labeled by specific human annotators, allowing the model to approximate annotator-specific labeling tendencies. Previous studies have explored ensemble approaches that train models on labels from different annotators to approximate multi-annotator supervision (Davani et al., 2022). LLMs leverage in-context learning without retraining, substantially reducing computational cost.

## 4 Soft Labeling Measurement

Empirically, the human annotation distribution $p_{\text{human}}(y \mid x)$ provides a natural reference point for evaluation, as the goal of LLM-based soft labeling (SL) is to capture empirical variation in human labeling judgments.

### 4.1 Individual-Level Agreement

Previous work typically evaluates LLM–human agreement using standard classification metrics (e.g., accuracy and F1) and intercoder agreement measures such as Cohen's Kappa or Krippendorff's Alpha (Basile et al., 2021; Vallejo Vera and Driggers, 2025). However, these metrics rely on *hard labels*, requiring annotations to be collapsed into a single label (i.e., majority votes or MV), making them incompatible with soft-label distributions.

In this study, we adopt an evaluation metric that can be applied consistently to both hard labels and soft labels, allowing for a direct comparison between human and LLM annotation distributions under both SL and MV settings. Specifically, we use mean absolute error (MAE) (Törnberg, 2024). In a binary annotation problem, MAE is defined as follows:

$$\mathrm{MAE} = \frac{1}{|D|} \sum_{i=1}^{|D|} \left| \ell_{\text{human}}(x_i) - \ell_{\text{LLM}}(x_i) \right|,$$

where $D$ denotes the set of items and $\ell(x)$ denotes the label representation used for comparison.

For *soft labels*, we use the probability distribution to calculate $\mathrm{MAE}_{SL}$:

$$\ell(x) = p(y \mid x),$$

For *majority-vote labels*, we convert the distribution into a hard label to calculate $\mathrm{MAE}_{MV}$:

$$\ell(x) = \mathbf{1}[p(y \mid x) > 0.5],$$

### 4.2 Downstream Performance

Downstream performance serves as a key criterion for directly comparing LLM-based annotation strategies under both MV and SL paradigms. MV typically uses F1 scores, while SL relies on cross-entropy (CE). To ensure consistency, we incorporate calibration metrics (e.g., Expected Calibration Error or ECE), which measure the alignment between predicted probabilities and empirical outcomes and are compatible with both settings.

**Traditional Metrics.** Traditional classification metrics such as accuracy, precision, recall, and F1 treat the gold label as a single deterministic category. These metrics apply directly to MV labels derived from the human annotation distribution. For models trained with SL, predicted class probabilities are converted to a single predicted label using the $\arg\max$ operator, allowing evaluation using the same metrics.

**Cross Entropy.** Recent work in NLP has proposed evaluating classification models using CE with SL, which reflects the full distribution of annotator judgments rather than enforcing a single correct label (Basile et al., 2021). For document $x_i$, CE is defined as:

$$\ell(x_i) = -\sum_{y \in \mathcal{Y}} p_{\text{human}}(y \mid x_i) \log p_{\text{LLM}}(y \mid x_i),$$

where $p_{\text{human}}(y \mid x_i)$ is the human soft-label distribution and $p_{\text{LLM}}(y \mid x_i)$ is the predicted probability distribution. Overall performance is computed as the mean CE across $N$ documents:

$$\mathrm{CE} = \frac{1}{|D|} \sum_{i=1}^{|D|} \ell(x_i).$$

**Calibration.** To evaluate the quality of the model's uncertainty estimates, we report ECE (Guo et al., 2017). ECE assesses how well predicted probabilities align with empirical outcomes, capturing model calibration under both MV and SL settings. Predictions are partitioned into $M$ confidence bins. Let $I_m$ denote the set of predictions whose confidence falls in bin $m$, ECE is defined as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|I_m|}{|D|} \left| \mathrm{acc}(I_m) - \mathrm{conf}(I_m) \right|.$$

We also compute a soft variant of ECE that accounts for annotation disagreement (Liang, 2026). For each instance, predicted confidence is defined as the maximum predicted probability, $\max_y p_{\text{LLM}}(y \mid x)$, while soft correctness is defined as the human probability assigned to the model's predicted class. $\mathrm{ECE}_{\text{soft}}$ is defined as:

$$\mathrm{ECE}_{\text{soft}} = \sum_{m=1}^{M} \frac{|I_m|}{D} \left| \mathrm{soft}(I_m) - \mathrm{conf}(I_m) \right|.$$

## 5 Experiments

### 5.1 Datasets

We evaluate our framework on the Gab Hate and GoEmotions datasets, both of which provide

annotator-level labels and anonymous IDs, enabling simulation of annotator-specific labeling behavior under *human-approximated variation*.

**Gab Hate (Target Label: Hate)** (Kennedy et al., 2022) is a dataset of 27,665 Gab posts (Gaffney, 2018) annotated for hate speech using a detailed codebook. Due to sparsity in other variables, we focus on the binary *hate* label (13% positive). Each post is labeled by at least three annotators from a pool of $|A| = 18$ (mean = 3.13 annotations per post). Following our experimental setup, we use an 80/20 train–test split and sample 10% of the training set for experiments.

**GoEmotions (Target Label: Admiration)** (Demszky et al., 2020) is a 58K Reddit emotion annotation benchmark with fine-grained multi-label emotion annotations. We focus on the *Admiration* label (8% positive), which exhibits relatively high inter-annotator reliability in the dataset. Each post is labeled by a pool of $|A| = 82$ annotators (mean = 3.64 annotations per post). We use the same 80/20 train–test split procedure and sample 5% of the training set for experiments.

For both datasets, we additionally sample $100 \times A$ annotator-specific demonstration examples from the remaining training data for in-context learning while ensuring no document overlap with training or test sets.

### 5.2 Model Selection

We evaluate a diverse set of closed- and open-source instruction-tuned models. For all LLM soft-labeling strategies (except model ensembles), we use GPT-4O-MINI and open-source models QWEN2.5-7B, LLAMA3-8B, MISTRAL-7B, and QWEN3-30B.

For the model ensemble strategy, we vary the ensemble composition to match the average annotator count of each dataset. For Gab Hate (3 annotators/document), we evaluate CLOSED-3 (GPT-4O-MINI, GEMINI-3-FLASH-PREVIEW, CLAUDE-HAIKU-4-5), OPEN-7B-3 (QWEN2.5-7B, LLAMA3-8B, MISTRAL-7B), 2O+1C (two open-source models and one closed-source model), and 2C+1O (two closed-source models and one open-source model).

For GoEmotions (4 annotators/document), we retain CLOSED-3 and OPEN-7B-3, and additionally evaluate two four-model hybrid ensembles: 2O+2C-1 and 2O+2C-2.

### 5.3 Experimental Setup

In this section, we describe the implementation details of the four soft-labeling strategies introduced in Section 3. To maintain comparability with the original human annotation structure, *model-induced variation* approaches generate LLM-based annotations that match the average number of annotations per document ($N$) in the benchmark datasets, whereas *human-structured variation* approaches simulate annotator pools ($A$) and preserve the original per-document annotation counts.

**Single-Model Repeated Sampling.** For each document, we independently query a single LLM $N$ times using stochastic decoding ($T = 0.7$). These annotations are treated as simulated annotator labels. Performance is evaluated separately for each model.

**Model Ensembles.** To introduce variation across model families, we construct several model ensembles. For each document, $N$ annotations are sampled from $N$ (i.e., three to four) different models within the ensemble using outputs generated in the single-model sampling stage. Annotations are sampled without replacement, and the ensemble construction is repeated with independent random samplings to account for randomness in ensemble composition. Performance for each ensemble configuration is averaged across runs.

**Demographic Persona Prompting.** To approximate structured variation observed in human annotation, we simulate a pool of $A$ annotators using demographic personas. Each persona is defined by a combination of demographic attributes sampled to roughly reflect distributions reported in the United States Census Bureau (2020). We consider three attributes associated with variation in labeling judgments: political ideology (PID $\in$ {conservative, liberal, moderate}), gender ({male, female}), and education level (EDU $\in$ {high school, bachelor's, graduate}). Three independent persona pools are generated to capture variation in demographic composition. Each persona generates one annotation per document using deterministic decoding ($T = 0$), as variation is introduced explicitly through persona conditioning. Results are averaged across the three persona pools.

**Human-Calibrated In-Context Annotation.** To approximate individual annotator behavior, we use in-context learning with examples labeled by specific human annotators. For each annotator, a subset of their previously labeled documents is in-

| Model | macro-F1 | CE | ECE | $ECE_{soft}$ |
|---|---|---|---|---|
| **Gab Hate - Hate** | | | | |
| Human Benchmark | | | | |
| Training Data | 0.052 | -0.138 | -0.060 | -0.036 |
| Model Ensembles | | | | |
| CLOSED-3 | 0.002 | 0.043 | -0.019 | 0.019 |
| OPEN-7B-3 | 0.001 | -0.179 | -0.061 | -0.039 |
| 2C+1O | -0.002 | -0.113 | -0.042 | -0.013 |
| 2O+1C | 0.014 | -0.133 | -0.056 | -0.037 |
| Human-Calibrated In-Context Annotation (50 Instances) | | | | |
| GPT-4O-MINI | 0.015 | -0.011 | -0.033 | -0.007 |
| LLAMA-3-8B-INSTRUCT | 0.004 | -0.124 | -0.042 | -0.026 |
| MISTRAL-7B-INSTRUCT | -0.001 | -0.041 | -0.016 | 0.010 |
| QWEN2.5-7B-INSTRUCT | 0.000 | 0.017 | -0.029 | 0.000 |
| QWEN3-30B-INSTRUCT | -0.001 | -0.019 | -0.025 | 0.000 |
| **GoEmotions - Admiration** | | | | |
| Human Benchmark | | | | |
| Training Data | -0.003 | -0.003 | -0.026 | -0.023 |
| Model Ensembles | | | | |
| CLOSED-3 | 0.007 | -0.033 | -0.019 | -0.002 |
| OPEN-7B-3 | 0.026 | -0.051 | -0.023 | -0.008 |
| 2C+2O-1 | 0.074 | -0.047 | -0.023 | -0.008 |
| 2C+2O-2 | 0.031 | -0.077 | -0.030 | -0.015 |
| Human-Calibrated In-Context Annotation (50 Instances) | | | | |
| GPT-4O-MINI | 0.067 | 0.003 | -0.006 | 0.011 |
| LLAMA-3-8B-INSTRUCT | 0.007 | -0.088 | -0.025 | -0.012 |
| MISTRAL-7B-INSTRUCT | 0.059 | -0.132 | -0.026 | -0.012 |
| QWEN2.5-7B-INSTRUCT | -0.030 | -0.039 | -0.013 | 0.000 |
| QWEN3-30B-INSTRUCT | 0.015 | -0.094 | -0.029 | -0.013 |

Table 1: **Soft Labeling (SL) consistently outperforms Majority Vote (MV) across most metrics.** Performance differences between SL and MV supervision ($\Delta$ = SL − MV) across the two datasets. Red values indicate performance degradation under SL.

cluded in the prompt as few-shot demonstrations, after which the model predicts labels for new documents. We construct four calibration settings with 10, 25, 50, and 100 examples per annotator, with examples randomized within the prompt to mitigate ordering effects. Each persona produces one annotation per document using deterministic decoding ($T = 0$), as variation is introduced explicitly through annotator-specific in-context examples. Performance is reported separately for each few-shot condition.

### 5.4 Evaluation

We evaluate LLM annotators under both SL and MV paradigms across two dimensions: (1) **Individual-level**: deviation from human annotations is measured using MAE on label distributions (SL) and, for comparison, MAE on hard labels derived via MV; (2) **Classification performance**: we fine-tune a ROBERTA classifier (AdamW, lr=1e-5, weight decay=0.01) with early stopping, averaged over five seeds. We report both SL and MV metrics, including macro-F1, ECE, $ECE_{soft}$, and CE (see Section 4 for details).

### 5.5 Results

**SL vs. MV in Subjective Tasks.** Table 1 reports the relative performance of SL over MV across evaluation metrics ($\Delta$ = SL − MV) for the two stronger strategies across the two datasets (weak ones are reported in Appendix A). SL yields consistent improvements in CE, with most configurations showing negative $\Delta CE$, indicating better alignment with the human label distribution.

SL also tends to improve calibration. Across both ensemble-based and human-calibrated in-context annotation settings, ECE and $ECE_{soft}$ generally decrease, indicating better correspondence between predicted confidence and empirical correctness. In contrast, improvements in macro-F1 are small and inconsistent, as expected for a metric based on hard predictions. Overall, these results suggest that SL is most effective when annotation strategies produce structured, informative disagreement rather than arbitrary variation.

**Comparison Across Four Strategies.** Table 2 (reporting average performance across model-specific results in Appendix B) compares annotation strategies under both MV and SL using human–LLM agreement and classification metrics. Overall, model ensembles achieve the strongest and most consistent performance across both datasets and all evaluation dimensions.

For the Gab Hate dataset, SL under model ensembles improves distributional alignment and calibration while maintaining comparable macro-F1. The same overall trend appears in GoEmotions, where SL consistently performs better on distribution-sensitive and calibration metrics. Ensemble methods also achieve relatively low MAE across datasets, indicating closer alignment with human annotation distributions.

Human-calibrated in-context annotation yields comparable results. Across both datasets, SL improves distributional and calibration metrics without sacrificing hard-label performance, suggesting that preserving annotator uncertainty helps models better approximate human label distributions.

More broadly, improvements under SL are more pronounced for CE and calibration metrics than for macro-F1 across annotation strategies. This pattern suggests that the primary advantage of SL lies in improving probabilistic alignment with human judgments and uncertainty estimation rather than substantially changing hard-label classification accuracy. Model-specific performance comparisons across annotation strategies for each dataset are reported in Appendix B.

**Human-Calibrated In-Context Annotation.** Figure 1 examines how performance varies with

| Condition | Majority Vote (MV) MAE ↓ | macro-F1 ↑ | CE ↓ | ECE ↓ | $ECE_{soft}$ ↓ | Soft Labeling (SL) MAE ↓ | macro-F1↑ | CE↓ | ECE ↓ | $ECE_{soft}$ ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Gab Hate - Hate** | | | | | | | | | | |
| **Human Benchmark** | – | 0.657 (0.046) | 0.526 (0.131) | 0.092 (0.017) | 0.092 (0.017) | – | 0.709 (0.021) | 0.388 (0.032) | 0.032 (0.010) | 0.056 (0.006) |
| Single-Model | **0.1016** | 0.656 (0.021) | 0.614 (0.229) | 0.114 (0.032) | 0.114 (0.032) | 0.1134 | 0.651 (0.025) | 0.706 (0.152) | 0.114 (0.038) | 0.141 (0.035) |
| Model Ensembles | **0.0820** | **0.674 (0.023)** | **0.575 (0.206)** | **0.104 (0.027)** | **0.104 (0.027)** | **0.0923** | **0.678 (0.020)** | **0.479 (0.066)** | **0.059 (0.017)** | **0.086 (0.016)** |
| Persona Prompting | 0.1260 | 0.640 (0.020) | 0.783 (0.280) | 0.147 (0.030) | 0.147 (0.030) | 0.1348 | **0.642 (0.021)** | 0.851 (0.200) | 0.140 (0.030) | 0.163 (0.027) |
| In-Context Annotation | 0.1082 | **0.658 (0.022)** | **0.564 (0.214)** | **0.112 (0.032)** | **0.112 (0.032)** | **0.1128** | 0.661 (0.022) | **0.528 (0.109)** | **0.083 (0.033)** | **0.107 (0.029)** |
| **GoEmotions - Admiration** | | | | | | | | | | |
| **Human Benchmark** | – | 0.794 (0.010) | 0.209 (0.019) | 0.041 (0.004) | 0.041 (0.004) | – | 0.791 (0.008) | 0.206 (0.011) | 0.015 (0.011) | 0.018 (0.006) |
| Single-Model | 0.0630 | **0.677 (0.037)** | **0.360 (0.096)** | **0.058 (0.011)** | **0.058 (0.011)** | **0.0780** | 0.699 (0.057) | 0.358 (0.078) | 0.045 (0.013) | 0.060 (0.013) |
| Model Ensembles | **0.0570** | **0.680 (0.071)** | **0.346 (0.072)** | **0.055 (0.010)** | **0.055 (0.010)** | **0.0720** | 0.715 (0.038) | **0.294 (0.036)** | **0.031 (0.007)** | **0.046 (0.008)** |
| Persona Prompting | **0.0620** | 0.657 (0.054) | 0.379 (0.087) | 0.060 (0.010) | 0.060 (0.010) | 0.0800 | **0.660 (0.058)** | 0.414 (0.081) | 0.055 (0.010) | 0.070 (0.010) |
| In-Context Annotation | 0.0700 | 0.659 (0.065) | 0.390 (0.129) | 0.060 (0.017) | 0.060 (0.017) | 0.0800 | **0.683 (0.056)** | **0.320 (0.032)** | **0.040 (0.010)** | **0.055 (0.010)** |

Table 2: **Model ensembles consistently achieved the best average performance across both datasets, followed by in-context annotation.** Aggregated performance comparison of annotation strategies under Majority Vote (MV) and Soft Labeling (SL) supervision for Gab Hate and GoEmotions. **Blue** bold indicates the best-performing result across all conditions, while **black** bold indicates the second-best result. In-context annotation uses 50 instances.

the number of annotator-specific examples in the in-context learning setup for Gab Hate. Overall, increasing the number of in-context annotations from 0 to 50 generally improves distribution-sensitive metrics (e.g. CE, ECE, $ECE_{soft}$). However, these gains are not monotonic. Performance typically peaks around 50 examples, after which additional context yields diminishing returns or slight degradation across several metrics. The right panel further shows that the advantage of SL over MV is strongest at moderate levels of in-context supervision, especially for CE.

Similar patterns are observed for GoEmotions (Figure 2 in Appendix C), where SL consistently improves CE and calibration relative to MV across most models. Unlike Gab Hate, however, GoEmotions also exhibits more substantial and consistent gains in macro-F1, suggesting that SL may provide stronger benefits for categorical prediction performance in emotion recognition tasks.

## 6 Discussion

Consistent with theoretical expectations and prior findings on human soft-label annotations (Collins et al., 2022; Fornaciari et al., 2021; Liang, 2026), SL training generally outperforms MV training across the evaluation spectrum (F1, ECE, $ECE_{soft}$, and CE) in our experiments on Gab Hate and GoEmotions, both of which involve substantial annotator disagreement (Table 1). Although certain metrics may occasionally favor MV supervision (e.g., F1 or ECE), SL training consistently provides stronger overall performance in settings that require modeling epistemic uncertainty arising from diverse human perspectives. These findings further motivate the use of LLM-based soft labeling as a scalable annotation strategy for tasks where such uncertainty is inherent.

Importantly, this advantage depends on whether the annotation strategy produces sufficiently informative disagreement signals. In particular, the SL advantage emerges for the two strongest LLM-based annotation strategies: model ensembles and annotator-calibrated in-context learning. In contrast, weaker variation mechanisms, such as repeated sampling from a single model or demographic persona prompting, do not reliably reproduce these benefits, as shown in Table 3. These patterns remain consistent across a broad range of models and across both the Gab Hate and GoEmotions datasets.

Recognizing these benefits, we formalize LLM-based soft labeling as introducing controlled variation in model-generated annotations through two conceptual sources: *model-induced variation* and *human-approximated variation*. This specification differs from many prior LLM-based annotation approaches that assume annotation diversity must arise from explicitly modeling human attributes (Fröhling et al., 2025).

Following this formulation, we observe that the most effective mechanism, model ensembles, does not require explicit representation of human attributes. Instead, it consistently outperform most single-model sampling and demographic persona prompting, with particularly strong performance under the SL training paradigm across all evaluation metrics and both datasets.

The second strongest strategy is human-calibrated in-context learning, which conditions models on examples labeled by individual annotators. Interestingly, in both datasets, performance peaks when approximately 50 annotated examples are provided. Increasing the number of examples beyond this point does not yield monotonic improvements and may even slightly degrade perfor-

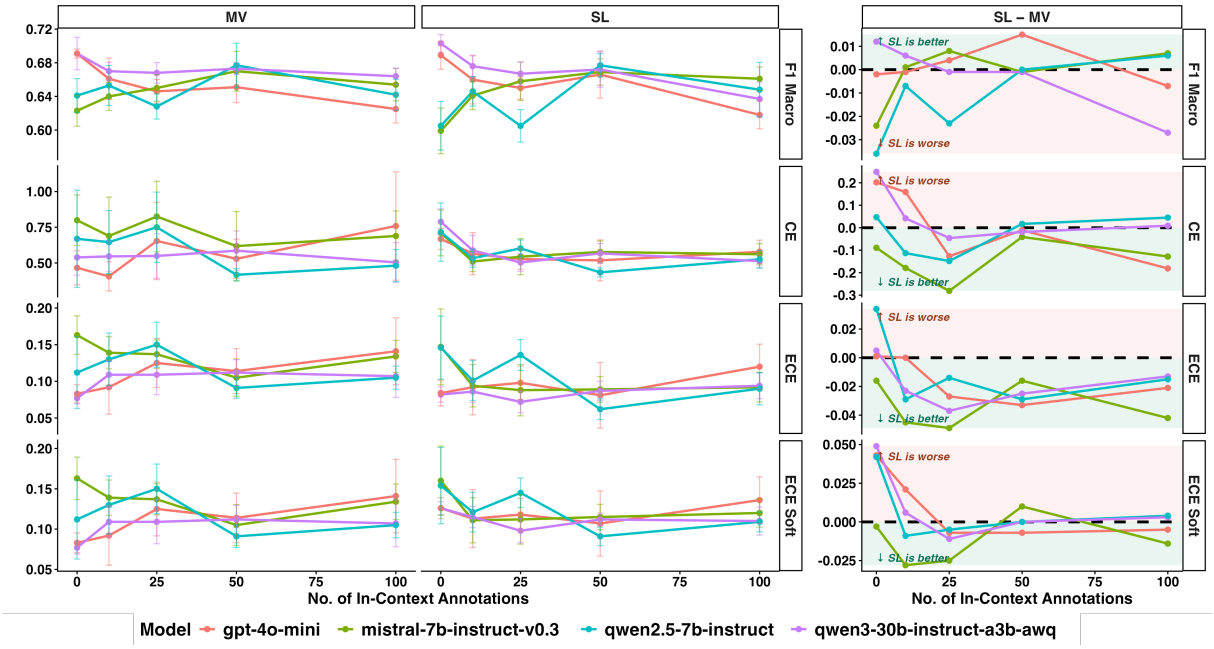


Figure 1: **Soft Labeling (SL) outperforms Majority Vote (MV) most strongly at moderate levels (25–50 instances) of in-context supervision.** Performance under MV and SL supervision across varying numbers of in-context annotations for Gab Hate. The right panel shows the difference ($\Delta$ = SL − MV). Error bars represent 95% confidence intervals across random seeds during RoBERTa training. Llama-3-8B-Instruct is omitted because its 8k-token context window cannot accommodate 100 instances.

mance. This pattern raises the possibility that excessive few-shot examples constrain model generalization or lead to overfitting to annotator-specific labeling patterns within the prompt (Tang et al., 2025). Investigating this behavior further may provide insight into the interaction between few-shot conditioning and annotator modeling.

Finally, single-model repeated sampling shows promising performance under the MV training paradigm, but performs less effectively under SL supervision. This pattern suggests that high-temperature sampling ($T = 0.7$) from a single model tends to reflect singular judgments rather than the heterogeneous perspectives required to produce informative soft-label distributions, aligning with previous findings on intra-model homogeneity (Jiang et al., 2025).

Across all strategies, these findings remain consistent across both individual-level human–LLM agreement and downstream task performance, and hold across a wide range of models as well as both hate speech and emotion annotation tasks.

## 7 Conclusion

We propose a formal perspective on LLM-based SL as introducing controlled sources of variation in model-generated annotations to approximate the latent variability present in human annotation processes. This formulation relaxes the assumption that such mechanisms must be explicitly designed to correspond to human factors in order to effectively utilize LLM-based SL.

Empirically, we find that model ensembles provide the most effective mechanism for generating informative soft-label distributions, consistent with our expectation. Variation across model training receipt (e.g., architectures, training data, and alignment procedures), can produce aggregated representational differences that generate useful disagreement signals, even without explicitly modeling human annotator characteristics. This observation also aligns with classical ensemble learning theory, where aggregating predictions from diverse models improves performance when individual models capture partially independent errors (Dietterich, 2000).

More broadly, our findings highlight an important design consideration for LLM-based annotation: mechanisms that induce independent predictive variation may be more effective than approaches that attempt to directly simulate human annotators, particularly given growing concerns about homogenization in LLM generation and annotation outputs (Jiang et al., 2025).

## Limitations

While this study evaluates both SL and MV training paradigms across a spectrum of evaluation metrics, we do not explicitly model the latent distributional differences between human and LLM annotators. Future work could more directly characterize these distributions by measuring divergences (e.g., KL divergence) or by parametrically approximating annotator pools using Dirichlet distributions. Such analyses could provide additional theoretical and empirical insight into how different LLM-based approaches inject variation into label distributions.

Our individual annotator calibration setting could also be extended using datasets that include richer annotator metadata, such as Spinde et al. (2021), which provides both annotator identities and demographic attributes. Incorporating both annotator behavioral history and demographic information may better approximate real human annotation behavior. Relatedly, future work could explore implicit persona prompting strategies that activate culturally associated cues (e.g., references to food, celebrities, or birthplace), which may influence model behavior without explicitly specifying demographic attributes (Wang et al., 2025).

Finally, although our experiments generalize across two target labels from the Gab Hate and GoEmotions datasets, both labels are binary variables with substantial class imbalance. Extending the analysis to additional annotation tasks, such as other emotions in GoEmotions or news bias labeling (Spinde et al., 2021), would help assess the generality of LLM-based soft labeling across domains with different label distributions and task characteristics.

## Ethics Statement

This study uses publicly available datasets released for research purposes and does not involve new data collection or personally identifiable information.

Following ACL responsible NLP guidelines, we note that tasks such as hate speech and stance classification involve inherently subjective judgments. To preserve this variation, we use soft labels rather than collapsing annotations into a single hard label.

We acknowledge that both the data and resulting models may reflect societal biases present in the underlying text and annotations. Model predictions should therefore be interpreted with caution, particularly in sensitive contexts. This work is intended for research purposes only and not for deployment in high-stakes or fully automated decision-making systems.

To support transparency and reproducibility, we report all model configurations, evaluation metrics, and experimental settings. Finally, improvements in predictive performance or calibration do not imply normative correctness, especially for subjective classification tasks.

## Acknowledgments

We appreciate the valuable feedback from the principal investigator, Dr. Elisa Kreiss, for insights on in-context learning, temperature sampling, and related work; Je Hoon Chae for suggestions on the formalization; and other students in the Computation and Language for Society (Coalas) Lab (https://www.coalas-lab.com/).

## References

Toufique Ahmed, Premkumar Devanbu, Christoph Treude, and Michael Pradel. 2025. Can LLMs Replace Manual Annotation of Software Engineering Artifacts? In *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*, pages 526–538. ISSN: 2574-3864.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D. Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

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

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a Large Language Model a Good Annotator for Event Extraction? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17772–17780.

Juhwan Choi, JungMin Yun, Kyohoon Jin, and Young-Bin Kim. 2024. Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15–29, Miami, Florida, USA. Association for Computational Linguistics.

Katherine M. Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and Learning with Soft Labels from Every Annotator. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10:40–52.

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

Sjoerd de Vries and Dirk Thierens. 2025. Learning with confidence: training better classifiers from soft labels. *Mach. Learn.*, 114(11).

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15.

Laura Espinosa and Marcel Salathé. 2024. Use of large language models as a scalable approach to understanding public health discourse. *PLOS Digital Health*, 3(10):e0000631.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2025. Personas with attitudes: Controlling LLMs for diverse data annotation. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 468–481, Vienna, Austria. Association for Computational Linguistics.

Gavin Gaffney. 2018. Pushshift gab corpus. `https://files.pushshift.io/gab/`. Accessed: 2019-05-23.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. LLMs Accelerate Annotation for Medical Information Extraction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 82–100. PMLR.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yanru Jiang and Gülşah Akçakır. 2025. Explicit cooperation shapes human-like multi-agent llm negotiation. In *Proceedings of the NLPSI 2025: First Workshop on Integrating NLP and Psychology to Study Social Interactions*, Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media (ICWSM 2025). Published: June 5, 2025.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, and 3 others. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Eval.*, 56(1):79–108.

Ayoung Lee, Ryan Sungmo Kwon, Peter Railton, and Lu Wang. 2026. CLASH: Evaluating language models on judging high-stakes dilemmas from multiple perspectives. In *The Fourteenth International Conference on Learning Representations*.

Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, Yimeng Zhang, Libo Qin, Yanyan Zhao, Qingfu Zhu, and Wanxiang Che. 2025. Can large language models understand you better? an MBTI personality detection dataset aligned with population traits. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5071–5081, Abu Dhabi, UAE. Association for Computational Linguistics.

Siyu Liang. 2026. Not Just Noise: Modeling Annotation Ambiguity in Text Analysis. Master's thesis, UCLA, February.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *arXiv preprint*. ArXiv:1908.07086 [cs].

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ranjan Sapkota, Achyut Paudel, and Manoj Karkee. 2025. Zero-Shot Automatic Annotation and Instance Segmentation using LLM-Generated Datasets: Eliminating Field Imaging and Manual Annotation for Deep Learning Model Development. *arXiv preprint*. ArXiv:2411.11285 [cs].

Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021. Mbic: A media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference 2021*, pages 1–8, Beijing, China (Virtual Event). Springer.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. *arXiv preprint*. ArXiv:2402.13446 [cs].

Yongjian Tang, Doruk Tuncel, Christian Körner, and Thomas Runkler. 2025. The few-shot dilemma: Overprompting large language models. *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 134–141.

Ramya Tekumalla and Juan M. Banda. 2023. Leveraging Large Language Models and Weak Supervision for Social Media data annotation: an evaluation using COVID-19 self-reported vaccination tweets. *arXiv preprint*. ArXiv:2309.06503 [cs].

Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. *arXiv preprint*. ArXiv:2402.05129 [cs].

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

United States Census Bureau. 2020. 2020 census demographic data. `https://data.census.gov`. United States Census Bureau.

Sebastián Vallejo Vera and Hunter Driggers. 2025. LLMs as annotators: the effect of party cues on labelling decisions by large language models. *Humanities and Social Sciences Communications*, 12(1):1530.

Lin Wang and Kuk-Jin Yoon. 2022. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068. ArXiv:2004.05937 [cs].

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual prompting for improving llm generation diversity. In *Conference on Empirical Methods in Natural Language Processing*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–21, New York, NY, USA. Association for Computing Machinery.

Peter Washington, Haik Kalantarian, Jack Kent, Arman Husic, Aaron Kline, Emilie Leblanc, Cathy Hou, Cezmi Mutlu, Kaitlyn Dunlap, Yordan Penev, Nate Stockham, Brianna Chrisman, Kelley Paskov, Jae-Yoon Jung, Catalin Voss, Nick Haber, and Dennis P. Wall. 2021. Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. *Cognitive computation*, 13(5):1363–1373.

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Jianfei Wu, Xubin Wang, and Weijia Jia. 2024. Enhancing Text Annotation through Rationale-Driven Collaborative Few-Shot Prompting. *arXiv preprint*. ArXiv:2409.09615 [cs].

Yuki Yamagata and Ryota Yamada. 2024. Survey on large language model annotation of cellular senescence from figures in review articles. *Genomics & Informatics*, 22:7.

Hua Yuan, Yu Shi, Ning Xu, Xu Yang, Xin Geng, and Yong Rui. 2023. Learning From Biased Soft Labels. *Advances in Neural Information Processing Systems*, 36:59566–59584.

| Model | macro-F1 | CE | ECE | $ECE_{soft}$ |
|---|---|---|---|---|
| **Gab Hate - Hate** | | | | |
| Human Benchmark | | | | |
| Training Data | 0.052 | -0.138 | -0.060 | -0.036 |
| Single-Model Repeated Sampling | | | | |
| GPT-4O-MINI | -0.002 | 0.202 | 0.001 | 0.043 |
| LLAMA-3-8B-INSTRUCT | 0.028 | 0.048 | -0.021 | 0.006 |
| MISTRAL-7B-INSTRUCT | -0.024 | -0.089 | -0.016 | -0.003 |
| QWEN2.5-7B-INSTRUCT | -0.036 | 0.047 | 0.034 | 0.042 |
| QWEN3-30B-INSTRUCT | 0.012 | 0.249 | 0.005 | 0.049 |
| Demographic Persona Prompting | | | | |
| GPT-4O-MINI | 0.003 | 0.169 | 0.004 | 0.035 |
| LLAMA-3-8B-INSTRUCT | -0.009 | 0.098 | -0.005 | 0.026 |
| MISTRAL-7B-INSTRUCT | 0.026 | -0.129 | -0.023 | -0.032 |
| QWEN2.5-7B-INSTRUCT | -0.005 | 0.035 | -0.008 | 0.013 |
| QWEN3-30B-INSTRUCT | -0.006 | 0.168 | -0.005 | 0.038 |
| **GoEmotions - Admiration** | | | | |
| Human Benchmark | | | | |
| Training Data | -0.003 | -0.003 | -0.026 | -0.023 |
| Single-Model Repeated Sampling | | | | |
| GPT-4O-MINI | 0.005 | -0.045 | -0.015 | 0.002 |
| LLAMA-3-8B-INSTRUCT | 0.141 | -0.045 | -0.021 | -0.007 |
| MISTRAL-7B-INSTRUCT | 0.019 | -0.006 | -0.011 | 0.003 |
| QWEN2.5-7B-INSTRUCT | -0.043 | -0.008 | -0.015 | 0.000 |
| QWEN3-30B-INSTRUCT | -0.010 | 0.097 | -0.003 | 0.014 |
| Demographic Persona Prompting | | | | |
| GPT-4O-MINI | -0.005 | 0.047 | -0.004 | 0.012 |
| LLAMA-3-8B-INSTRUCT | 0.032 | 0.051 | -0.005 | 0.009 |
| MISTRAL-7B-INSTRUCT | -0.014 | 0.044 | -0.002 | 0.011 |
| QWEN2.5-7B-INSTRUCT | -0.004 | 0.026 | -0.006 | 0.010 |
| QWEN3-30B-INSTRUCT | 0.002 | 0.010 | -0.008 | 0.008 |

Table 3: **Performance difference between SL and MV supervision ($\Delta$ = SL − MV) for the weaker strategies across the two datasets.** Red values indicate performance degradation under SL.

## A SL vs. MV in Weaker Strategies

Table 3 reports the performance differences between SL and MV under the two additional weaker annotation strategies. Under these weaker supervision settings, MV remains comparatively competitive, likely because it provides a more stable training signal.

## B Model-Specific Performance

Tables 4 and 5 report model-specific performance comparisons across annotation strategies for each dataset, respectively.

For the Gab Hate dataset, model ensemble SL substantially improves distributional alignment, reducing CE from 0.575 to 0.479 and improving calibration, with ECE decreasing from 0.104 to 0.059 and $ECE_{soft}$ decreasing from 0.104 to 0.086, while maintaining comparable macro-F1. A similar pattern emerges for GoEmotions, where CE decreases from 0.346 to 0.294 and ECE decreases from 0.055 to 0.031 under SL. Ensemble methods also achieve relatively low MAE across datasets (Gab Hate: MV = 0.082, SL = 0.092; GoEmotions: MV = 0.057, SL = 0.072), indicating closer alignment with human annotation distributions.

Human-calibrated in-context annotation exhibits a similar trend. On Gab Hate, SL reduces CE from 0.564 to 0.528 and improves calibration (ECE: 0.112 to 0.083; $ECE_{soft}$: 0.112 to 0.107), while maintaining stable macro-F1 (0.658 vs. 0.661). On GoEmotions, CE decreases from 0.390 to 0.320, with ECE improving from 0.060 to 0.040 and $ECE_{soft}$ decreasing from 0.060 to 0.055.

In contrast, weaker annotation strategies perform less reliably under SL. For single-model sampling on Gab Hate, SL increases CE from 0.614 to 0.706 and worsens calibration ($ECE_{soft}$: 0.114 to 0.141). A similar but smaller degradation appears on GoEmotions, where $ECE_{soft}$ increases from 0.058 to 0.060. Persona prompting exhibits the same pattern, with higher CE under SL for both Gab Hate (0.783 to 0.851) and GoEmotions (0.379 to 0.414), alongside worse calibration (Gab Hate $ECE_{soft}$: 0.147 to 0.163; GoEmotions: 0.060 to 0.070). These results suggest that demographic variation alone does not necessarily produce informative or well-calibrated label distributions.

## C Human-Calibrated In-Context Annotation of GoEmotions

Figure 2 illustrates human-calibrated in-context annotation on GoEmotions. Patterns are consistent with those observed for the Gab Hate dataset, where SL consistently improves CE and calibration relative to MV across most models. In contrast to Gab Hate, however, GoEmotions also exhibits larger and more consistent gains in macro-F1.

| Model | Majority Vote (MV) MAE ↓ | macro-F1 ↑ | CE ↓ | ECE ↓ | $ECE_{soft}$ ↓ | Soft Labeling (SL) MAE ↓ | macro-F1 ↑ | CE ↓ | ECE ↓ | $ECE_{soft}$ ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Human Benchmark** | | | | | | | | | | |
| | | 0.657 (0.046) | 0.526 (0.131) | 0.092 (0.017) | 0.092 (0.017) | | 0.709 (0.021) | 0.388 (0.032) | 0.032 (0.010) | 0.056 (0.006) |
| **Single-Model Repeated Sampling** | | | | | | | | | | |
| GPT-4O-MINI | **0.059** | **0.691 (0.006)** | **0.466 (0.136)** | 0.083 (0.014) | 0.083 (0.014) | **0.08** | 0.689 (0.019) | 0.668 (0.084) | 0.084 (0.020) | **0.126 (0.015)** |
| LLAMA-3-8B-INSTRUCT | 0.098 | 0.632 (0.025) | 0.598 (0.178) | 0.133 (0.027) | 0.133 (0.027) | 0.106 | 0.660 (0.026) | **0.646 (0.102)** | 0.112 (0.026) | 0.139 (0.020) |
| MISTRAL-7B-INSTRUCT | 0.194 | 0.623 (0.021) | 0.799 (0.202) | 0.163 (0.030) | 0.163 (0.030) | 0.19 | 0.599 (0.031) | 0.710 (0.182) | 0.147 (0.059) | 0.160 (0.049) |
| QWEN2.5-7B-INSTRUCT | 0.098 | 0.641 (0.023) | 0.669 (0.388) | 0.112 (0.056) | 0.112 (0.056) | 0.109 | 0.605 (0.033) | 0.716 (0.233) | 0.146 (0.049) | 0.154 (0.054) |
| QWEN3-30B-INSTRUCT | **0.059** | **0.691 (0.022)** | 0.539 (0.142) | **0.077 (0.009)** | **0.077 (0.009)** | 0.082 | **0.703 (0.012)** | 0.788 (0.104) | **0.082 (0.012)** | **0.126 (0.009)** |
| Avg. | 0.102 | 0.656 (0.021) | 0.614 (0.229) | 0.114 (0.032) | 0.114 (0.032) | 0.113 | 0.651 (0.025) | 0.706 (0.152) | 0.114 (0.038) | 0.141 (0.035) |
| **Model Ensembles** | | | | | | | | | | |
| CLOSED-3 | **0.060** | **0.694 (0.020)** | **0.467 (0.133)** | **0.082 (0.021)** | **0.082 (0.021)** | **0.072** | **0.696 (0.016)** | 0.510 (0.061) | 0.063 (0.013) | 0.101 (0.014) |
| OPEN-7B-3 | 0.107 | 0.661 (0.024) | 0.677 (0.270) | 0.123 (0.030) | 0.123 (0.030) | 0.116 | 0.662 (0.018) | 0.498 (0.078) | 0.062 (0.019) | 0.084 (0.017) |
| 2C+1O | 0.063 | 0.693 (0.023) | 0.545 (0.175) | 0.092 (0.015) | 0.092 (0.015) | 0.074 | 0.691 (0.024) | **0.432 (0.063)** | **0.050 (0.020)** | **0.079 (0.019)** |
| 2O+1C | 0.098 | 0.650 (0.025) | 0.610 (0.220) | 0.117 (0.038) | 0.117 (0.038) | 0.107 | 0.664 (0.020) | 0.477 (0.059) | 0.061 (0.014) | 0.080 (0.015) |
| Avg. | **0.082** | **0.674 (0.023)** | 0.575 (0.206) | **0.104 (0.027)** | **0.104 (0.027)** | **0.092** | **0.678 (0.020)** | **0.479 (0.066)** | **0.059 (0.017)** | **0.086 (0.016)** |
| **Demographic Persona Prompting** | | | | | | | | | | |
| GPT-4O-MINI | 0.081 | 0.676 (0.024) | **0.532 (0.184)** | 0.099 (0.030) | 0.099 (0.030) | 0.093 | **0.679 (0.021)** | **0.701 (0.123)** | 0.103 (0.026) | 0.134 (0.020) |
| LLAMA-3-8B-INSTRUCT | 0.085 | 0.679 (0.013) | 0.659 (0.140) | 0.108 (0.015) | 0.108 (0.015) | 0.099 | 0.670 (0.016) | 0.757 (0.206) | 0.103 (0.032) | 0.134 (0.033) |
| MISTRAL-7B-INSTRUCT | 0.305 | 0.494 (0.020) | 1.427 (0.537) | 0.311 (0.045) | 0.311 (0.045) | 0.29 | 0.520 (0.029) | 1.298 (0.295) | 0.288 (0.041) | 0.288 (0.041) |
| QWEN2.5-7B-INSTRUCT | 0.103 | 0.658 (0.022) | 0.744 (0.213) | 0.133 (0.033) | 0.133 (0.033) | 0.115 | 0.653 (0.016) | 0.779 (0.214) | 0.125 (0.032) | 0.146 (0.031) |
| QWEN3-30B-INSTRUCT | **0.056** | **0.692 (0.019)** | 0.551 (0.071) | **0.085 (0.011)** | **0.085 (0.011)** | **0.077** | 0.686 (0.019) | 0.719 (0.097) | **0.080 (0.011)** | **0.123 (0.009)** |
| Avg. | 0.126 | 0.640 (0.020) | 0.783 (0.280) | 0.147 (0.030) | 0.147 (0.030) | 0.135 | 0.642 (0.021) | 0.851 (0.200) | 0.140 (0.030) | 0.163 (0.027) |
| **Human-Calibrated In-Context Annotation (50 Instances)** | | | | | | | | | | |
| GPT-4O-MINI | 0.117 | 0.651 (0.021) | 0.529 (0.157) | 0.114 (0.035) | 0.114 (0.035) | 0.119 | 0.666 (0.032) | 0.518 (0.164) | 0.081 (0.051) | 0.107 (0.046) |
| LLAMA-3-8B-INSTRUCT | 0.127 | 0.618 (0.019) | 0.670 (0.318) | 0.137 (0.048) | 0.137 (0.048) | 0.129 | 0.622 (0.012) | 0.546 (0.131) | 0.095 (0.044) | 0.111 (0.040) |
| MISTRAL-7B-INSTRUCT | 0.095 | 0.670 (0.027) | 0.618 (0.276) | 0.105 (0.029) | 0.105 (0.029) | 0.104 | 0.669 (0.018) | 0.577 (0.088) | 0.089 (0.020) | 0.115 (0.018) |
| QWEN2.5-7B-INSTRUCT | **0.094** | **0.677 (0.030)** | **0.417 (0.050)** | **0.091 (0.016)** | **0.091 (0.016)** | **0.101** | **0.677 (0.016)** | **0.434 (0.036)** | **0.062 (0.016)** | **0.091 (0.013)** |
| QWEN3-30B-INSTRUCT | 0.108 | 0.673 (0.008) | 0.586 (0.158) | 0.112 (0.020) | 0.112 (0.020) | 0.111 | 0.672 (0.024) | 0.567 (0.079) | 0.087 (0.015) | 0.112 (0.009) |
| Avg. | 0.108 | 0.658 (0.022) | **0.564 (0.214)** | 0.112 (0.032) | 0.112 (0.032) | 0.113 | 0.661 (0.022) | 0.528 (0.109) | 0.083 (0.033) | 0.107 (0.029) |

Table 4: **Gab Hate performance comparison across annotation strategies under MV and SL. Bold** denotes the best within each group; **blue** marks the overall best across models and strategies; **green** indicates the best strategy by group average, consistent with the aggregated results reported in Table 2.

| Model | Majority Vote (MV) MAE ↓ | macro-F1 ↑ | CE ↓ | ECE ↓ | $ECE_{soft}$ ↓ | Soft Labeling (SL) MAE ↓ | macro-F1 ↑ | CE ↓ | ECE ↓ | $ECE_{soft}$ ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Human Benchmark** | | | | | | | | | | |
| | | 0.794 (0.010) | 0.209 (0.019) | 0.041 (0.004) | 0.041 (0.004) | | 0.791 (0.008) | 0.206 (0.011) | 0.015 (0.011) | 0.018 (0.006) |
| **Single-Model Repeated Sampling** | | | | | | | | | | |
| GPT-4O-MINI | **0.053** | **0.744 (0.014)** | 0.381 (0.035) | 0.056 (0.001) | 0.056 (0.001) | **0.069** | **0.749 (0.016)** | **0.336 (0.059)** | **0.041 (0.007)** | **0.058 (0.008)** |
| LLAMA-3-8B-INSTRUCT | 0.067 | 0.511 (0.067) | 0.391 (0.114) | 0.065 (0.011) | 0.065 (0.011) | 0.078 | 0.652 (0.039) | 0.346 (0.047) | 0.044 (0.010) | **0.058 (0.009)** |
| MISTRAL-7B-INSTRUCT | 0.065 | 0.672 (0.040) | 0.372 (0.120) | 0.057 (0.017) | 0.057 (0.017) | 0.082 | 0.691 (0.026) | 0.366 (0.086) | 0.046 (0.020) | 0.060 (0.019) |
| QWEN2.5-7B-INSTRUCT | 0.068 | 0.729 (0.018) | 0.349 (0.093) | 0.059 (0.010) | 0.059 (0.010) | 0.083 | 0.686 (0.115) | 0.341 (0.060) | 0.044 (0.014) | 0.059 (0.015) |
| QWEN3-30B-INSTRUCT | 0.061 | 0.729 (0.011) | **0.305 (0.096)** | **0.051 (0.012)** | **0.051 (0.012)** | 0.079 | 0.719 (0.018) | 0.402 (0.117) | 0.048 (0.012) | 0.065 (0.012) |
| Avg. | 0.063 | 0.677 (0.037) | 0.360 (0.096) | 0.058 (0.011) | 0.058 (0.011) | 0.078 | 0.699 (0.057) | 0.358 (0.078) | 0.045 (0.013) | 0.060 (0.013) |
| **Model Ensembles** | | | | | | | | | | |
| CLOSED-3 | **0.055** | **0.746 (0.017)** | **0.321 (0.080)** | **0.050 (0.012)** | **0.050 (0.012)** | 0.071 | **0.753 (0.011)** | 0.288 (0.033) | 0.031 (0.007) | 0.048 (0.006) |
| OPEN-7B-3 | 0.061 | 0.661 (0.080) | 0.361 (0.068) | 0.056 (0.013) | 0.056 (0.013) | 0.074 | 0.687 (0.048) | 0.310 (0.040) | 0.033 (0.007) | 0.048 (0.008) |
| 2C2O-1 | **0.055** | 0.632 (0.098) | 0.342 (0.075) | 0.054 (0.008) | 0.054 (0.008) | **0.069** | 0.706 (0.043) | 0.295 (0.033) | 0.031 (0.006) | 0.046 (0.007) |
| 2O2C-2 | 0.058 | 0.682 (0.060) | 0.360 (0.062) | 0.059 (0.005) | 0.059 (0.005) | 0.073 | 0.713 (0.037) | **0.283 (0.039)** | **0.029 (0.009)** | **0.044 (0.011)** |
| Avg. | **0.057** | **0.680 (0.071)** | **0.346 (0.072)** | **0.055 (0.010)** | **0.055 (0.010)** | **0.072** | **0.715 (0.038)** | **0.294 (0.036)** | **0.031 (0.007)** | **0.046 (0.008)** |
| **Demographic Persona Prompting** | | | | | | | | | | |
| GPT-4O-MINI | **0.054** | **0.719 (0.029)** | **0.342 (0.093)** | **0.053 (0.007)** | **0.053 (0.007)** | **0.073** | 0.714 (0.056) | **0.389 (0.054)** | **0.049 (0.005)** | **0.065 (0.006)** |
| LLAMA-3-8B-INSTRUCT | 0.071 | 0.574 (0.080) | 0.400 (0.090) | 0.066 (0.014) | 0.066 (0.014) | 0.087 | 0.606 (0.074) | 0.451 (0.100) | 0.061 (0.014) | 0.075 (0.014) |
| MISTRAL-7B-INSTRUCT | 0.062 | 0.565 (0.082) | 0.389 (0.070) | 0.063 (0.008) | 0.063 (0.008) | 0.080 | 0.551 (0.085) | 0.433 (0.075) | 0.061 (0.005) | 0.074 (0.005) |
| QWEN2.5-7B-INSTRUCT | 0.063 | 0.714 (0.018) | 0.381 (0.088) | 0.058 (0.006) | 0.058 (0.006) | 0.082 | 0.710 (0.018) | 0.407 (0.083) | 0.052 (0.009) | 0.068 (0.009) |
| QWEN3-30B-INSTRUCT | 0.061 | 0.715 (0.019) | 0.382 (0.093) | 0.058 (0.012) | 0.058 (0.012) | 0.078 | **0.717 (0.023)** | 0.392 (0.086) | 0.050 (0.014) | 0.066 (0.014) |
| Avg. | 0.062 | 0.657 (0.054) | 0.379 (0.087) | 0.060 (0.010) | 0.060 (0.010) | 0.080 | 0.660 (0.058) | 0.414 (0.081) | 0.055 (0.010) | 0.070 (0.010) |
| **Human-Calibrated In-Context Annotation (50 Instances)** | | | | | | | | | | |
| GPT-4O-MINI | **0.058** | 0.680 (0.112) | **0.325 (0.097)** | **0.046 (0.020)** | **0.046 (0.020)** | **0.071** | **0.747 (0.023)** | 0.328 (0.012) | 0.040 (0.006) | 0.057 (0.005) |
| LLAMA-3-8B-INSTRUCT | 0.080 | 0.675 (0.006) | 0.386 (0.138) | 0.059 (0.024) | 0.059 (0.024) | 0.090 | 0.682 (0.014) | 0.298 (0.033) | 0.034 (0.013) | 0.047 (0.013) |
| MISTRAL-7B-INSTRUCT | 0.074 | 0.573 (0.085) | 0.465 (0.119) | 0.072 (0.011) | 0.072 (0.011) | 0.083 | 0.632 (0.087) | 0.333 (0.031) | 0.046 (0.009) | 0.060 (0.009) |
| QWEN2.5-7B-INSTRUCT | 0.070 | 0.667 (0.022) | 0.408 (0.144) | 0.066 (0.014) | 0.066 (0.014) | 0.082 | 0.637 (0.087) | 0.369 (0.045) | 0.053 (0.010) | 0.066 (0.009) |
| QWEN3-30B-INSTRUCT | 0.067 | **0.702 (0.025)** | 0.365 (0.142) | 0.056 (0.013) | 0.056 (0.013) | 0.076 | 0.717 (0.008) | **0.271 (0.031)** | **0.027 (0.010)** | **0.043 (0.011)** |
| Avg. | 0.070 | 0.659 (0.065) | 0.390 (0.129) | 0.060 (0.017) | 0.060 (0.017) | 0.080 | 0.683 (0.056) | 0.320 (0.032) | 0.040 (0.010) | 0.055 (0.010) |

Table 5: **GoEmotions performance comparison across annotation strategies under MV and SL. Bold** denotes the best within each group; **blue** marks the overall best across models and strategies; **green** indicates the best strategy by group average, consistent with the aggregated results reported in Table 2.

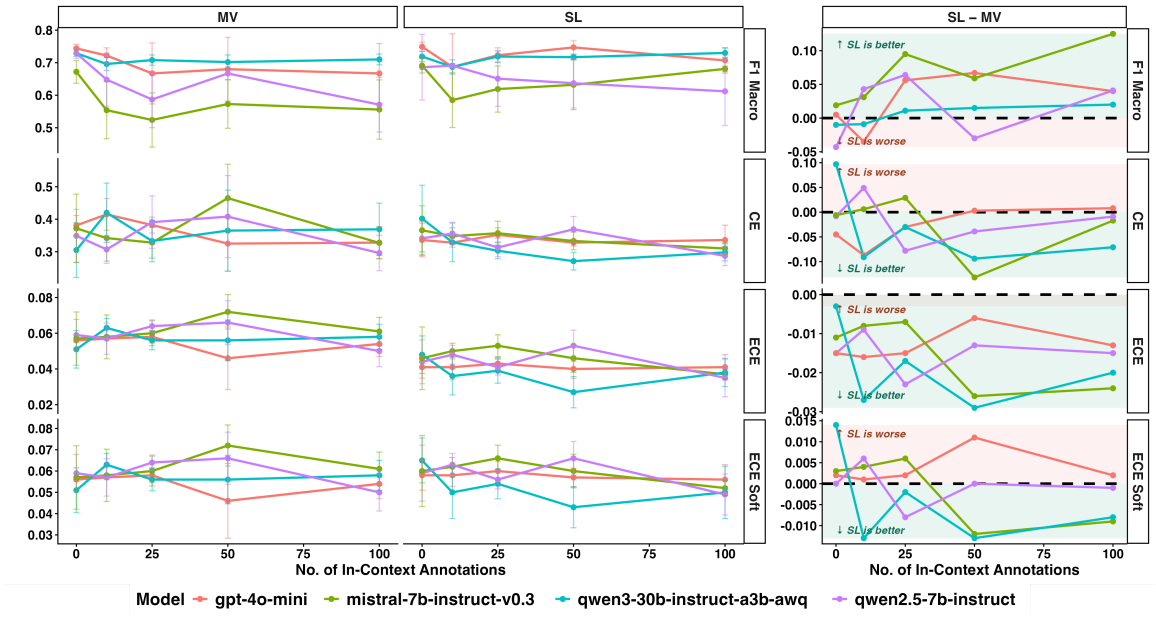


Figure 2: **Performance under MV and SL supervision across varying numbers of in-context annotations for GoEmotions.** The right panel shows the difference ($\Delta = \text{SL} - \text{MV}$). Error bars represent 95% confidence intervals across random seeds in RoBERTa training. LLAMA-3-8B-INSTRUCT is omitted because its 8k-token context window cannot accommodate 100 instances.