Using Model Disagreement to Identify Unstable Regions in MT Evaluation

Vitalii Iakivchuk
Smartling
244 Fifth Avenue Suite 1471 New York, NY 10001
viakivchuk@smartling.com

Abstract

Human evaluation of MT is essential but exhibits substantial annotator variability that limits evaluation reliability and supervised learning. Rather than treating disagreement as noise or correcting it through protocol changes, we analyze its structure via learned severity classifiers.

Across training regimes defined by baseline model reproducibility, we observe internally coherent but mutually incompatible severity mappings: models trained on one regime produce confident predictions within that regime but reduced separability on the other. Margin–correctness analysis shows that instability is not uniformly low confidence; separability depends on alignment between model-internalized and human annotation regimes.

These results indicate that unstable MT evaluation regions are primarily associated with competing severity interpretations rather than intrinsic example difficulty. Model–annotator disagreement therefore provides a practical signal for identifying unstable evaluation regions during MT evaluation.

1 Introduction

Human MT evaluation remains the primary assessment standard but shows substantial inter-annotator variability (Popović, 2021; Plank, 2022; Song et al., 2025), limiting both evaluation reliability and the usefulness of labeled data (Tan and Monz, 2025).

Prior work has often attributed inter-annotator disagreement in MT evaluation to annotation difficulty or unclear guidelines (Lommel et al., 2014; Popović, 2021) and has tried to mitigate it through refined protocols (Lommel et al., 2014; Popović, 2021) or AI assistance (Ni et al., 2026; Zouhar et al., 2025). However, such disagreement can also arise from legitimate variability in how annotators interpret error severity (Popović, 2021; Basile et al., 2021; Xu et al., 2026).

Our approach examines whether evaluation data supports a transferable severity mapping across different subsets of the data. We study how predictive behavior changes when severity classifiers are trained on different regions of the dataset, and whether learned decision patterns extend consistently beyond their training region. We partition a large-scale multilingual LQA dataset into subsets defined by baseline model reproducibility and train separate classifiers within each subset.

We further analyze prediction separability and confidence characteristics across these regimes in order to understand how disagreement manifests in model behavior. In particular, we examine whether regions associated with reduced transferability are characterized by uniformly low certainty or instead by shifts in the relative preference between competing severity alternatives. This analytical lens allows us to explore the possibility that instability in MT evaluation reflects differences in annotation structure and interpretation rather than solely intrinsic example difficulty or random variability.

Taken together, this perspective motivates the use of model–annotator disagreement as a practical signal for identifying evaluation regions that may require additional analysis or consistency support

during annotation.

1.1 Contributions

- We show that MT evaluation data can contain internally coherent but mutually incompatible severity mappings.
- We demonstrate regime-dependent separability: model confidence depends on alignment between training subset and evaluation region, where different regions may have similar confidence but different correctness.
- We provide evidence that unstable evaluation regions are not uniformly low-confidence and are consistent with competing annotation interpretations.
- We propose model–annotator disagreement as a deployable signal for identifying unstable regions during MT evaluation.

2 Related Work

Prior work typically attributes disagreement in MT evaluation to annotation difficulty, insufficient guidance, and inherent variability in severity interpretation (Lommel et al., 2014; Popović, 2021; Freitag et al., 2021). Such inconsistency also affects downstream metric learning, where noisy human ratings complicate the training of evaluation models (Tan and Monz, 2025; Moghe et al., 2025). These challenges echo broader annotation issues such as label ambiguity and task subjectivity observed across NLP tasks (Pavlick and Kwiatkowski, 2019; Klie et al., 2022). Proposed solutions range from protocol refinements and AI-assisted workflows (Zouhar et al., 2025) to approaches that model disagreement as meaningful signal rather than noise (Xu et al., 2026; Fleisig et al., 2023; Ni et al., 2026).

Protocol refinements focus on clearer guidelines and better annotation practices. Freitag et al. (2021) conduct a large-scale MQM study and show that crowd-worker evaluations without document context yield substantially different rankings from expert MQM annotations, proposing the use of professional translators for more reliable evaluations. Similarly, Lommel et al. (2014) examine inter-annotator agreement in error annotation, identifying factors like ambiguous categorizations, span disagreements, and varying severity perceptions as sources of difficulty, and suggest protocol improvements including better instructions and decision-making tools. Popović (2021) analyzes disagreements in MT output evaluation, noting that unclear quality criteria contribute to variability, and recommends refining definitions to enhance agreement, while acknowledging inherent challenges in complex linguistic structures such as negation or relative clauses.

AI-assisted workflows aim to reduce cognitive load while preserving quality. Zouhar et al. (2025) note that automated metrics remain misaligned with human judgment and propose AI-assisted annotation by pre-filling error spans to reduce time and maintain quality, finding that the unified priming also increases inter-annotator agreement.

At the same time, a growing body of evidence shows that disagreement is not merely noise or the result of poor guidance. It can reflect genuine variability in how annotators interpret severity. Fleisig et al. (2023) model annotator disagreement in subjective tasks like hate speech detection, arguing it stems from systematic demographic differences rather than noise, and demonstrate that majority votes can overlook minority perspectives, such as those of targeted groups. Similarly, Davani et al. (2022) propose multi-annotator models that preserve individual annotator judgments as separate subtasks with shared representations, showing that this approach matches or outperforms majority-vote aggregation across multiple subjective classification tasks and yields uncertainty estimates that correlate with annotator disagreement. Freitag et al. (2021) observe that professional annotations reveal a clear preference for human over machine translations that crowd-worker evaluations fail to capture, suggesting that different evaluator populations perceive quality differently. Popović (2021) finds that for phenomena spanning multiple words or phrases, such as negation or relative clauses, disagreements are inherent to the evaluation process rather than errors. Xu et al. (2026) argue that human label variation should be treated not as noise to eliminate but as an intrinsic value reflecting pluralistic human perspectives, and call for preserving annotator-level labels in preference datasets. Ni et al. (2026) investigate whether LLM reasoning can capture human annotator disagreement, finding that chain-of-thought prompting with RLHF models improves disagreement modeling while RLVR-style reasoning degrades it.

In a broader NLP context, Klie et al. (2022)

analyze annotation error detection across diverse datasets and find that ambiguous instances, where multiple labels are valid given the context, are prevalent and harder to detect than outright errors.

In contrast to prior approaches, which primarily aim to minimize or statistically model disagreement, our work analyzes the internal structure of disagreement and demonstrates that translation quality evaluation data behave as if they contain competing latent severity interpretations.

3 Method

3.1 Dataset, Severity Definition, and Experimental Structure

We use a large-scale LQA dataset of 34,192 machine-translated segments across 20+ target locales and 18 error categories from professional localization workflows. Each instance consists of a source segment and its translation, enriched with source-aligned glossary terms and contextual metadata (target locale, formality level, and translation domain). Segments are annotated by professional evaluators using full MQM-style annotations, including error categories and severity levels. In this study, we focus on the segment-level severity label, defined as the maximum severity among errors identified in the segment.

To ensure sufficient representation of both correct and erroneous translations for analysis, the dataset was constructed to contain approximately balanced proportions of segments with and without annotated errors.

All severity classifiers in this study (Models A, B, and C) are fine-tuned instances of the GPT-4o large language model trained to predict segment severity. All three models use identical fine-tuning configuration and differ only in their training data. Model confidence scores are derived from token-level log-probabilities returned during output generation.

To ensure sufficient support for each label, we merge adjacent severity levels into a three-level scale:

- None — no error or neutral
- Minor — minor
- Major — major or critical

Severity distinctions follow MQM-style evaluation principles, where *Minor* errors typically affect local fluency or limited adequacy issues, while *Major* errors impact meaning preservation or translation usability.

To maintain statistical power, all languages are pooled and analyzed jointly (no language-specific stratification).

Experimental structure (single overview). We first split the full dataset D into three disjoint partitions:

- D_{train} — used to train the baseline classifier
- D_{test} — used to construct the reproducibility regimes
- D_{hold} — external holdout used only for evaluation

$D = D_{\text{train}} \cup D_{\text{test}} \cup D_{\text{hold}}$, and all three partitions are mutually disjoint.

Reproducibility regimes (constructed on D_{test}). We train a baseline severity classifier (Model A) on D_{train}. We then run Model A on D_{test} and partition D_{test} into:

- S — segments in D_{test} correctly predicted by Model A
- U — segments in D_{test} mispredicted by Model A

$D_{\text{test}} = S \cup U$, where S and U are disjoint. The partition is defined empirically based on agreement or disagreement between Model A predictions and the reference annotation.

Regime-specific training splits. We further split each regime into disjoint training and holdout portions:

- $S_{\text{train}}, S_{\text{hold}}$, with $S = S_{\text{train}} \cup S_{\text{hold}}$
- $U_{\text{train}}, U_{\text{hold}}$, with $U = U_{\text{train}} \cup U_{\text{hold}}$

All four splits are subsets of D_{test} and are disjoint from both D_{train} and D_{hold}.

Models and evaluation sets. We train Model B on S_{train} and Model C on U_{train}. We evaluate Models B and C on:

- S_{hold}
- U_{hold}
- D_{hold}

This design enables comparison of regime-specific learnability and cross-regime behavior, while keeping D_{hold} as an external evaluation set not used in regime construction.

3.2 Margin-Based Confidence

We further examine each model's prediction confidence using the log-probability margin between the top two severity predictions. We extract the log-probabilities returned by the GPT-4o API at the output generation step, taking the values corresponding to the three severity tokens (*None*, *Minor*, *Major*) and treating them as class scores $s(y) = \log P(y)$. The margin is defined as

$$\text{margin} = s(y_1) - s(y_2),$$

where y_1 and y_2 are the severity labels with the highest and second-highest scores. A high margin indicates a strong preference for one severity label over competing alternatives, while a low margin indicates weaker separation between candidate labels.

Large margins can occur when the model assigns high probability to a single label. For example, if the model assigns probability 0.99 to *Major* and 0.00005 to the next most likely label (e.g., *Minor*), then $\log(0.99) \approx -0.01$ and $\log(0.00005) \approx -10$, yielding a margin of approximately 10. No additional normalization is applied.

3.3 Experimental Schema

Figure 1 summarizes the data partitioning and model training pipeline described above.

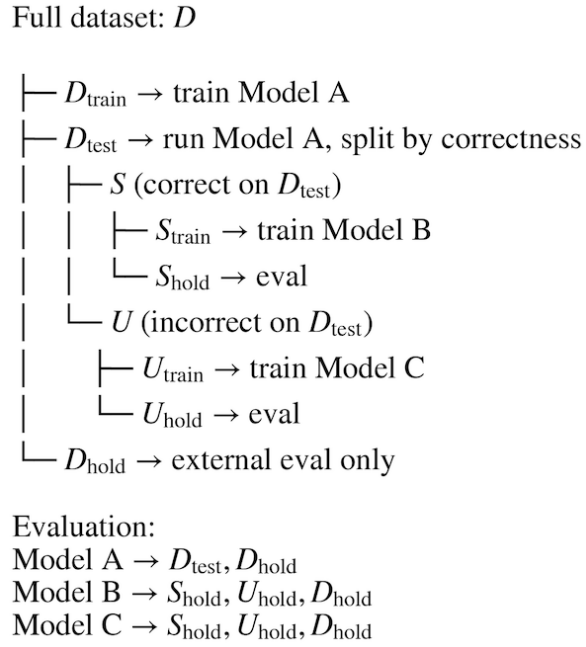

Figure 1: Experimental schema overview.

3.4 Training and Evaluation Procedure

The experimental procedure consists of five steps:

1. Split D into D_{train} ($\sim$30%, including a development subset D_{dev} used for validation), D_{test} ($\sim$55%), and D_{hold} ($\sim$15%). Fine-tune GPT-4o on D_{train}, yielding Model A.
2. Run Model A on D_{test} and partition into S (correct predictions) and U (incorrect predictions).
3. Split each regime into training, development, and holdout portions (approximately 70%/15%/15%): S_{train}, S_{dev}, S_{hold} and U_{train}, U_{dev}, U_{hold}.
4. Fine-tune two additional models using identical configuration: Model B on S_{train} (with S_{dev} for validation), Model C on U_{train} (with U_{dev}).
5. Evaluate all models on S_{hold}, U_{hold}, and D_{hold}, extracting token-level log-probabilities for margin analysis.

4 Empirical Analysis of Regime Structure

This section analyzes predictive performance and separability behavior across reproducibility regimes. We examine whether subsets defined by Model A correctness exhibit distinct learnability and confidence structure.

Both regimes span diverse locales and error types. Table 1 shows that major locales appear in both regimes.

Both S and U contain all 18 error categories present in the dataset, with the same five most frequent categories (no error, mistranslation, unidiomatic, inconsistency, omission) in both.

4.1 Regime-Specific Predictive Performance

We first compare predictive performance across evaluation subsets using three-class severity (*None* / *Minor* / *Major*). Table 2 summarizes results for all models across S_{hold}, U_{hold}, and D_{hold}.

Three observations emerge:

- **Stable regime coherence.** Model B trained on S achieves very high performance on S_{hold} (0.913 accuracy).
- **Cross-regime incompatibility.** The same model collapses on U_{hold} (0.208), while Model C shows the opposite pattern.

Table 1: Locale distribution across regimes (top 15 by total count).

Locale	S	U	% in U
de-DE	779	457	37.0
fr-FR	722	511	41.4
zh-TW	681	517	43.2
pt-PT	805	366	31.3
it-IT	675	458	40.4
fr-CA	649	467	41.8
pt-BR	686	425	38.3
nl-NL	672	398	37.2
ja-JP	620	442	41.6
ko	554	373	40.2
fi-FI	558	349	38.5
sv-SE	516	323	38.5
zh-CN	488	282	36.6
es-ES	386	307	44.3
pl-PL	388	284	42.3

- **No benefit of U for global generalization.** Model B approximately matches Model A on D_{hold} (0.625 vs 0.617), indicating that U does not improve overall predictive mapping.

This behavior is not driven solely by the majority *None* class. Macro F1 scores in Table 2 indicate that performance degradation affects all severity levels rather than a single dominant class.

Bootstrap confidence intervals confirm that in-regime and cross-regime accuracy do not overlap: Model B achieves 0.913 [0.899, 0.926] on S_{hold} versus 0.208 [0.184, 0.232] on U_{hold}, and Model C achieves 0.534 [0.505, 0.562] on U_{hold} versus 0.152 [0.135, 0.169] on S_{hold} (Table 3).

These results are consistent with S and U containing internally consistent but mutually incompatible severity mappings.

4.2 Margin–Correctness Relationship

We next examine whether margin-based confidence predicts correctness. We evaluate accuracy at several predefined margin thresholds chosen to illustrate the relationship between confidence and correctness. For a given threshold t, accuracy is computed on predictions with margin $\geq t$, while coverage denotes the proportion of predictions satisfying this condition.

On D_{hold}, Model B's margin moderately predicts correctness (AUC = 0.647, n = 5092). Increasing the minimum required margin improves accuracy on the retained predictions while reducing coverage (Table 4).

Within the stable regime, Model B's margin aligns strongly with correctness: on S_{hold}, AUC = 0.730 (n = 1672). Accuracy increases rapidly with threshold (Table 5).

Overall, margin reflects separability within regimes but does not distinguish regime identity.

4.3 Regime-Dependent Separability

We analyze margin distributions across models and regimes (Figure 2).

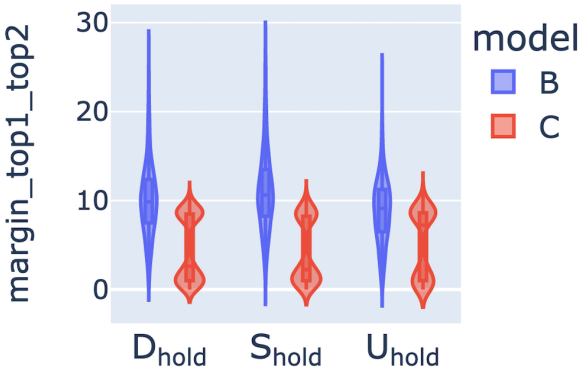

Figure 2: Violin plots of margin distributions across regimes.

Key observations.

- Model B shows consistently high margins across regimes.
- Model C exhibits a bimodal margin structure.

Model B shows unimodal margin distributions across holdouts; combined with its high accuracy on S_{hold}, this indicates that S forms a coherent (effectively unimodal) regime. The bimodal margins of Model C suggest heterogeneous structure in U (Figure 2).

4.4 Margin–Correctness Structure by Regime

We further examine margin conditioned on correctness (Figure 3).

Patterns.

- For Model B, margin distributions for correct and incorrect predictions are broadly similar, indicating limited global alignment between confidence and correctness.
- For Model C, particularly on U_{hold}, higher margins are often associated with incorrect predictions.

Table 2: Predictive performance across regimes.

Model	Training	Evaluation	Accuracy	Macro F1	n
A	D_{train}	D_{hold}	0.617	0.568	5130
B	S_{train}	S_{hold}	0.913	0.893	1695
B	S_{train}	U_{hold}	0.208	0.204	1128
B	S_{train}	D_{hold}	0.625	0.577	5130
C	U_{train}	U_{hold}	0.534	0.516	1128
C	U_{train}	S_{hold}	0.152	0.179	1695
C	U_{train}	D_{hold}	0.283	0.286	5130

Table 3: Bootstrap 95% confidence intervals for cross-regime accuracy.

Condition	Accuracy	CI Low	CI High
B on S_{hold}	0.913	0.899	0.926
B on U_{hold}	0.208	0.184	0.232
C on S_{hold}	0.152	0.135	0.169
C on U_{hold}	0.534	0.505	0.562

Table 4: Model B threshold effects on D_{hold}.

Threshold	Coverage	Accuracy
0	1.00	0.625
5	0.91	0.645
8	0.70	0.678
10	0.49	0.722

Table 5: Model B threshold effects on S_{hold}.

Threshold	Coverage	Accuracy
0	1.00	0.913
5	0.94	0.922
8	0.76	0.939
10	0.57	0.958

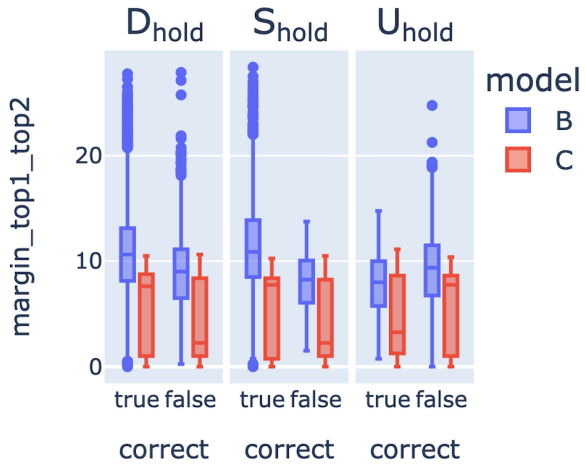

Figure 3: Box plots of margin distributions by correctness and regime.

- On U_{hold}, Model B exhibits comparatively higher margins for incorrect than for correct predictions, indicating high-confidence decisions that are misaligned with the reference annotations.

Thus, margin reflects ambiguity within a severity mapping rather than regime identity (Figure 3).

4.5 Structure of Unstable Regions

Combining predictive and margin analyses:

- S is highly coherent and transferable.
- U is partially learnable but incompatible with S.
- Margin does not uniformly collapse on U.
- U contains high-confidence predictions under a different mapping.

This is consistent with unstable regions reflecting competing severity interpretations rather than intrinsic difficulty.

4.6 Summary of Empirical Findings

The empirical analysis suggests:

- MT evaluation data can exhibit internally coherent but incompatible regimes.
- Predictive performance depends on regime alignment.
- Margin reflects confidence within regimes, not regime membership.
- Unstable regions are consistent with conflicting annotation mappings.

5 Discussion

The observed separability asymmetry is consistent with S and U corresponding to mutually incompatible yet internally coherent severity mappings. However, because these subsets are defined solely by baseline correctness, this partition alone does not guarantee that S should exhibit consistent or learnable structure, particularly under noisy human annotation where evaluators may disagree with themselves and with others (Popović, 2021). Selection by correctness therefore only increases the likelihood that consistent patterns, if present, appear in S rather than in U.

A related concern is whether the regime structure is solely an artefact of Model A's internal self-consistency. However, the cross-regime failure is symmetric: Model B collapses on U_{hold} (0.208) and Model C collapses on S_{hold} (0.152). All three models use the same architecture (GPT-4o) and identical fine-tuning procedure, differing only in training data. If the partition merely reflected Model A's bias, Model C would be expected to reproduce a compatible mapping on S. Instead, Model C learns a distinct mapping that is incompatible with S, suggesting that the two regimes reflect differences in the underlying annotations rather than a single model's bias.

Empirically, Model B trained on S_{train} achieves strong performance on S_{hold}, consistent with stable labeling structure in S. In contrast, Model C trained on U_{train} attains only moderate performance on U_{hold}, suggesting that U is more heterogeneous.

If instability were driven primarily by example difficulty or missing context, all models would exhibit uniformly reduced separability on U. Instead, separability depends on alignment between the training subset and the evaluation regime: models remain confident within their training regime while degrading sharply outside it. This pattern is consistent with disagreement reflecting competing annotation mappings rather than intrinsic ambiguity of the underlying segments.

These findings challenge approaches that treat annotator disagreement purely as noise to be minimized and instead support a perspectivist view of annotation in which variability reflects structured differences in interpretation (Xu et al., 2026). Prior work has shown that disagreement in MT evaluation is partly inherent due to subjectivity (Popović, 2021) and may reflect population-level differences in judgment (Fleisig et al., 2023; Ni et al., 2026). Our results extend this perspective by showing that MT evaluation data can contain distinct severity regimes whose incompatibility produces persistent low agreement.

Practically, regime structure suggests that disagreement can be used diagnostically. Regime-aware models could identify segments likely to receive inconsistent severity labels and prioritize them for targeted review. More broadly, incorporating regime information or annotator-conditioned modeling may improve robustness of MT evaluation datasets and downstream training signals.

6 Conclusion

We show that unstable regions in MT evaluation are largely associated with competing severity interpretations rather than solely with intrinsic example difficulty. While prior work highlights additional contributors such as annotation noise, protocol variation, and contextual uncertainty, our results indicate that evaluation data may contain internally coherent yet mutually incompatible severity mappings, giving rise to regime-dependent separability across trained models.

Because instability often appears as regime mismatch rather than uniformly low confidence, model–annotator disagreement provides a practical indicator of unstable evaluation regions. This reframes model assistance in MT evaluation from confidence-based filtering toward regime-aware disagreement detection, enabling targeted consistency support without assuming model correctness. Embracing structured annotation variability may therefore improve both MT evaluation reliability and supervision quality.

7 Limitations

The findings of this study should be interpreted within the scope of the experimental design.

First, reproducibility regimes are defined operationally through the predictive behavior of a single baseline severity classifier. The resulting stable and unstable subsets therefore correspond to regions of the dataset that exhibit higher or lower predictive reproducibility under the present modeling configuration. While this enables controlled analysis of regime-dependent transferability, the partition should not be interpreted as a definitive

identification of latent annotation regimes independent of modeling assumptions.

Second, the analysis is conducted at the segment level using a single aggregated severity label per segment. This representation abstracts away span-level structure, multiple concurrent errors, and annotator-level disagreement distributions. Consequently, the observed regime structure reflects patterns in segment-level severity prediction rather than the richer structure of MQM-style annotation decisions.

Third, all language pairs are pooled in order to maintain sufficient statistical power for regime-specific training and evaluation. The reported effects therefore characterize aggregate behavior across languages and may obscure language-specific variation in regime structure or severity interpretation.

Finally, the study provides observational evidence about predictive transferability and separability patterns across regimes. While the results are consistent with the presence of competing severity interpretations, the experimental design does not directly identify the cognitive or procedural sources of such disagreement.

These limitations do not compromise the internal consistency of the reported empirical patterns but constrain the extent to which broader causal or theoretical conclusions can be drawn.

Sustainability Statement

This study involves computational experiments using the proprietary GPT-4o model accessed through managed external API infrastructure. The experiments were conducted on a dataset of 34,192 machine-translated segments.

Training compute. Three severity classification models were fine-tuned once each with early stopping enabled, processing approximately 19.6 million training tokens in total.

Inference compute. Model inference was required for regime construction on D_{test} and for evaluation on all holdout subsets for the respective models. Overall, inference processing in this study consumed approximately 12 million input tokens.

Because the underlying hardware configuration, energy efficiency characteristics, and data center energy sources are controlled by the API provider, precise estimates of energy consumption or carbon emissions are not available. We therefore report dataset scale, training token counts, and inference workload characteristics to improve transparency regarding computational resource usage.

References

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In Church, Kenneth, Mark Liberman, and Valia Kordoni, editors, *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online, August. Association for Computational Linguistics.

Fleisig, Eve, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore, December. Association for Computational Linguistics.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Klie, Jan-Christoph, Bonnie Webber, and Iryna Gurevych. 2022. Annotation error detection: Analyzing the past and present for a more coherent future.

Lommel, Arle, Maja Popović, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE), LREC 2014*, pages 31–37, Reykjavik, Iceland, May.

Moghe, Nikita, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137, March.

Mostafazadeh Davani, Aida, 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.

Ni, Jingwei, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2026. Can reasoning help large language models capture human annotator disagreement?

Pavlick, Ellie and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

Popović, Maja. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In Bisazza, Arianna and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online, November. Association for Computational Linguistics.

Song, Yixiao, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing human evaluation in machine translation with comparative judgement. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20536–20551, Vienna, Austria, July. Association for Computational Linguistics.

Tan, Shaomu and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling.

Xu, Shanshan, Santosh T. Y. S. S, and Barbara Plank. 2026. From noise to signal to selbstzweck: Reframing human label variation in the era of post-training in nlp.

Zouhar, Vilém, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico, April. Association for Computational Linguistics.