Beyond Averages: Learning with Annotator Disagreement in STS

Alejandro Benito-Santos and Adrián Ghajari
UNED NLP&IR Group
Department of Languages and Systems
Universidad Nacional de Educación a Distancia (UNED)
Correspondence: al.benito@lsi.uned.es

Abstract

This work investigates capturing and modeling disagreement in Semantic Textual Similarity (STS), where sentence pairs are assigned ordinal similarity labels (0–5). Conventional STS systems average multiple annotator scores and focus on a single numeric estimate, overlooking label dispersion. By leveraging the disaggregated SemEval-2015 dataset (Soft-STS-15), this paper proposes and compares two disagreement-aware strategies that treat STS as an ordinal distribution prediction problem: a lightweight truncated Gaussian head for standard regression models, and a cross-encoder trained with a distance-aware objective, refined with temperature scaling. Results show improved performance in distance-based metrics, with the calibrated soft-label model proving best overall and notably more accurate on the most ambiguous pairs. This demonstrates that modeling disagreement benefits both calibration and ranking accuracy, highlighting the value of retaining and modeling full annotation distributions rather than collapsing them to a single mean label.

1 Introduction

Semantic Textual Similarity (STS) (Cer et al., 2017) is a core task in Natural Language Processing, used in applications such as question answering, information retrieval (Reimers et al., 2016), retrieval-augmented generation (RAG) (Lewis et al., 2020), or text de-duplication. Typically, annotators score sentence pairs on a six-point scale, but benchmark evaluations (e.g., STS-B) average these multiple labels into one mean value. This consolidation overlooks inherent label dispersion and incentivizes models to ignore the uncertainty crucial for practical settings (Basile, 2021). Real-world uses, however, rarely treat similarity as a strict binary decision; in fact, crowd and expert annotations often disagree around moderate similarity levels, suggesting systematic rather than random disagreement (Knupleš et al., 2023).

In this work, we investigate the benefits of modeling the full distribution of annotator labels. To this end, we structure our investigation around a logical progression: First, we ask whether capturing this distribution improves model *calibration* to human judgments, especially when respecting the empirical ordinal distances in the STS scale (**RQ1**). Recognizing that many downstream tasks still require scalar scores, we then examine whether this improved calibration comes at the cost of standard STS utility metrics like correlation or RMSE (**RQ2**). Finally, we explore the trade-offs between training with a soft-label objective versus applying purely post-hoc calibration to a standard regression-trained model (**RQ3**). To answer these questions, our experiments [1] compare a baseline cross-encoder (fine-tuned on STS-B) against two training paradigms (hard vs. soft-label learning).

Evaluations on both scalar and distributional metrics show that modeling the full annotator distribution significantly reduces calibration error to human label variation while simultaneously improving traditional metrics like rank correlation and accuracy on the central tendency (RMSE).

2 Related Work

Our work sits at the intersection of Semantic Textual Similarity (STS) and the growing field of Learning with Disagreements (LeWiDi).

Learning with Disagreements. There is a growing recognition in NLP that disagreement among annotators is often not noise, but a signal reflecting inherent ambiguity or subjectivity (Basile, 2021; Uma et al., 2020). This trend is captured by the "Learning-With-Disagreements" (LeWiDi) paradigm, which advocates for modeling the full

[1] **Code & data:** https://github.com/ale0xb/sts_beyond_averages

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35540–35546

distribution of human annotations rather than collapsing them to a single ground truth. This perspective is crucial for tasks involving ordinal scales, where recent analysis shows that mid-scale averages often conflate distinct, systematic patterns of human judgment (Knupleš et al., 2023).

Disagreement and Uncertainty in STS. While STS benchmarks typically rely on averaged scores (Cer et al., 2017), the importance of uncertainty and calibration in this regression setting is gaining attention: For example, recently Wang et al. (2022) focused on quantifying the calibration of pre-trained models for text regression.

More directly related to our study, Wang et al. (2023) analyzed collective human opinions in STS, introducing a large-scale Chinese dataset with disaggregated labels. They demonstrated that current models trained on averaged labels often fail to capture the variance caused by human disagreement. Our work builds on this foundation by focusing specifically on *ordinal* calibration using the English Soft-STS-15 dataset. To this end, we introduce a distance-aware strategy that incorporates empirically-derived perceptual distances into both the training objective (OLL) and the post-hoc calibration process. To the best of our knowledge, ours is the first approach to test this approach on the STS task.

3 Data

We use the original dataset collected via Amazon Mechanical Turk for SemEval-2015 (Agirre et al., 2015). The dataset comprises 8,387 sentence pairs annotated for semantic equivalence from various sources, which reduced to 7,890 pairs after excluding control questions. The data was retrieved from the original website of the SemEval STS tasks [2]. To our knowledge, this represents the only publicly available disaggregated version of data collected across the SemEval tasks conducted from 2012 to 2016. In this paper we refer to this dataset simply as "Soft-STS-15".

3.1 Modeling Disagreement

To quantify the level of disagreement for each sentence pair while respecting the ordinal nature of the STS scale, we utilize Krippendorff's alpha (α) (Krippendorff, 2011). This metric measures agreement relative to what would be expected by chance, allowing us to distinguish systematic disagreement from random noise. Specifically, we compute Krippendorff's (Krippendorff, 2011) per-item agreement score, α_i, for each item $i \in \{1, \ldots, N\}$ in the Soft-STS-15 dataset. This score provides finer-grained insight into items that may exhibit higher or lower levels of coder disagreement which is the basis of our study. This metric leverages the globally computed distance matrix D (Equation 4) and the global expected (random chance) coincidence matrix E, derived from the marginal frequencies n_v across all items: Let E_{jk} be the entry in E for the label pair (l_j, l_k). The global expected disagreement rate, $\bar{d}_E$, is calculated as the average distance expected by chance:

$$\bar{d}_E = \frac{\sum_{j,k} E_{jk} D_{jk}}{\sum_{j,k} E_{jk}}, \tag{1}$$

assuming $\sum_{j,k} E_{jk} > 0$. For each individual item i, we compute its specific observed coincidence matrix O_i, where $(O_i)_{jk}$ represents the observed frequency of the label pair (l_j, l_k) among coders for that item, normalized by the number of pairable ratings for item i. The observed disagreement rate for item i, $\bar{d}_{O,i}$, measures the actual average distance between the labels assigned by annotators for that item:

$$\bar{d}_{O,i} = \frac{\sum_{j,k} (O_i)_{jk} D_{jk}}{\sum_{j,k} (O_i)_{jk}}, \tag{2}$$

assuming $\sum_{j,k} (O_i)_{jk} > 0$. The per-item alpha agreement score, α_i, compares the item's observed disagreement rate to the global expected disagreement rate:

$$\alpha_i = 1 - \frac{\bar{d}_{O,i}}{\bar{d}_E}, \tag{3}$$

provided $\bar{d}_E$ is non-zero. An α_i close to 1 indicates that the observed disagreement for item i is much lower than expected by chance globally, while lower or negative values suggest higher-than-expected disagreement for that specific item.

We leverage a ground distance matrix $D \in \mathbb{R}^{V \times V}$, derived empirically using Krippendorff's alpha framework with an ordinal level of measurement. Specifically, let n_k be the marginal frequency (total count) of label l_k observed in the full reliability dataset used for Krippendorff's analysis. The distance d_{ij} between labels l_i and l_j (assuming $i \leq j$ without loss of generality due to symmetry)

[2] http://ixa2.si.ehu.es/stswiki/images/2/21/STS2015-en-rawdata-scripts.zip

is calculated as:

$$d_{ij} = \left(\sum_{k=i}^{j} n_k - \frac{n_i + n_j}{2} \right)^2 \quad (4)$$

From these distances, we empirically construct the distance matrix D using the cumulative frequency of intervening labels as:

$$D = \left(d(C_i, C_j) \right)_{(i,j) \in [\![1,6]\!]^2} \quad (5)$$

where $C = (C_1 \dots C_6)$ correspond to the STS annotation labels $\{0, 1, \dots 5\}$. A depiction of matrix D is shown in Figure 1.

3.2 Data Splits

Following a similar approach to (Leonardelli et al., 2021), we split the data into three different levels of per-item agreement (terciles) holding low, moderate, and high agreement pairs, respectively. The data is further split by majority label (i.e. the mode of the human annotations) to obtain a 7×3 $\{0, 1, \dots 5, \varnothing\} \times \{T1, T2, T3\}$ grid (majority label $\times$ tercile) [3]. A contingency table of the resulting 21 strata can be consulted in Appendix B. Finally, we partition the 7 890 pairs into a 60:20:20 split (4 734/1 578/1 578 items) *stratified* over the Cartesian product of majority labels (MODE $\in \{-1, 0, \dots, 5\}$) and disagreement tercile ($q \in \{1, 2, 3\}$). Stratification guarantees that even rare slices such as $\langle 5, \mathrm{T1} \rangle$ are represented in every partition, preventing skew towards majoritarian labels in the evaluation sets and enabling fair slice-wise calibration analysis.

4 Training

To measure the effects of soft label learning, we fine-tuned multiple `RoBERTa-large` cross-encoders using the `sentence-transformers` library (Reimers and Gurevych, 2019). We selected this architecture as the basis for our experiments because it represents a strong, state-of-the-art approach for STS; specifically, it is utilized by the top-performing models in the `sentence-transformers` library, ensuring a comparison against standard best practices. Using this architecture, we compare two learning approaches:

Soft-label learning. The cross-encoder is fine-tuned on disaggregated labels from STS-15 by utilizing a modified version of the cross-entropy loss, known as the ordinal log-loss (OLL) (Castagnos et al., 2022). This particular loss function was selected due to its demonstrated effectiveness in the original study on ordinal text classification tasks akin to STS, outperforming other established alternatives such as EMD, CORAL, or Cross-Entropy (Uma et al., 2020). Future studies will explore the impact of alternative loss functions within our framework.

$$\mathcal{L}_{\text{OLL-}\alpha}(P, y) = -\sum_{i=1}^{N} \log(1 - p_i)\, d(y, i)^{\alpha} \quad (6)$$

In our experiments, we use Krippendorff's distance matrix from Equation 5. To avoid contamination during training, we build the distance matrix D on the label distribution of the training and development sets of subsection 3.2. In our tests, we found setting the hyperparameter $\alpha = 1.5$ yielded the best results, in line with previous findings by Castagnos et al. (2022).

Hard label learning. In a hard learning configuration, we train the system to regress the mean of annotation scores in STS-15 employing a binary cross-entropy (BCE) loss.

For both objectives, we keep the encoder backbone, optimiser, and schedule *identical* so that the only experimental variable is the treatment of label uncertainty. All cross-encoders are initialized from the public checkpoint of `roberta-large`[4] and optimised with AdamW ($\eta = 1 \times 10^{-5}$, and weight-decay 0.01). Training proceeds for four epochs with mini-batches of 64 examples, a linear warm-up of ten percent of the total steps, and no gradient clipping. We apply early stopping by monitoring either δ-EMD on the development split for the soft objective or Pearson correlation of the mean score for the regression objective every 80 update steps; training is halted after three consecutive non-improving checkpoints. All experiments are repeated with five reproducible seeds and results are averaged. Runs were executed on a single NVIDIA 4090 RTX GPU taking approximately 2 hours to complete.

5 Evaluation

The trained cross-encoders are both evaluated on a regression and a soft scenario. Regression on STS datasets (Soft-STS-15 and STS-B) is measured using Spearman´s rank correlation, which should be

[3] We use $\varnothing$ to denote the category of items without a mode.

[4] https://huggingface.co/FacebookAI/roberta-large

the preferred metric (Reimers et al., 2016). The average prediction error in the regression task is captured by the root mean squared error (RMSE). For soft evaluation, we rely on a metric aimed at ordinal quantification, Earth-Mover's Distance (EMD), following recommendations by Sakai (2021). Intuitively, EMD measures the minimum "work" required to transform the predicted probability distribution into the gold distribution, where the cost of moving probability mass depends on the distance between the ordinal bins. Below, EMD is defined:

Let $\mathbf{y}, \hat{\mathbf{p}} \in \Delta^{K-1}$ be the gold and predicted distributions on the ordinal label set $\mathcal{L} = \{0, \ldots, K-1\}$ and $\mathbf{\Delta} = (\Delta_{jk})$ the Krippendorff disagreement matrix introduced in subsection 3.1. We define EMD as:

$$\begin{aligned} \mathrm{EMD}_{\mathbf{\Delta}}(\mathbf{y}, \hat{\mathbf{p}}) &= \min_{\mathbf{T} \in \mathbb{R}_{\geq 0}^{K \times K}} \sum_{j,k} \Delta_{jk}\, T_{jk} \\ \text{subject to} \quad & \mathbf{T}\mathbf{1} = \mathbf{y}, \quad \mathbf{T}^{\top}\mathbf{1} = \hat{\mathbf{p}} \end{aligned} \tag{7}$$

EMD quantifies the minimum transport cost needed to move the probability mass of $\hat{\mathbf{p}}$ onto $\mathbf{y}$ when moving one unit of mass from bin j to k costs Δ_{jk}. As we note in Figure 1, the perceptual distance between categories is not uniform across the STS ordinal levels of the scale. Therefore, we incorporate this notion into the EMD metric by using Krippendorff's distance matrix to compute the transport costs between ordinal labels (i.e., $\Delta_{ij} = D_{ij}$). Subsequently, we refer to this distance-aware EMD as δ-EMD.

Because we evaluate soft-label models with hard (scalar) metrics and hard-label models with soft (distributional) metrics, we require bidirectional mappings—one that transforms ordinal probability vectors into single regression scores, and another that converts scalar predictions back into ordinal distributions. Both procedures are explained hereafter.

Regression → Ordinal To transform a point estimate $\hat{m} \in [0, 1]$ from the regression head into a probability distribution over the six ordinal bins we place a *truncated Gaussian kernel* centred at the rescaled location $\hat{m}(K-1) \in [0, 5]$:

$$\mathrm{TN}_k(\hat{m}, \sigma) = \frac{\exp\!\left(-\frac{(k - \hat{m}(K-1))^2}{2\sigma^2}\right)}{\sum_{j=0}^{5} \exp\!\left(-\frac{(j - \hat{m}(K-1))^2}{2\sigma^2}\right)}$$

To provide a realistic reconstruction of the original variance, instead of a single global bandwidth we learn a separate variance $\sigma_{(u,q)}$ for every *mode–tercile*, effectively calibrating the output distribution to human uncertainty. Each $\sigma_{(u,q)}$ is chosen to minimise the distance-aware Earth–Mover loss on the development data:

$$\sigma^{\star}_{(u,q)} = \arg\min_{\sigma > 0} \; \delta\text{-EMD}\big(\mathrm{TN}(\hat{m}, \sigma), \mathbf{y}_{\text{dev}}\big).$$

At inference time the slice-specific $\sigma^{\star}_{(u,q)}$ is applied to the item, producing a calibrated yet computationally light uncertainty head that requires no additional gradient updates. The resulting distributions capture the dispersion observed among annotators and therefore provide a faithful, slice-aware approximation of label uncertainty.

Ordinal → Regression We enable the inverse operation for the soft-trained models when a single point estimate is required in the evaluation (i.e., for Spearman/RMSE). To do this, we convert the probability vector $\hat{\mathbf{p}} = (\hat{p}_0, \ldots, \hat{p}_5)$ predicted by the soft model into a scalar by taking its *expected label* and normalising to the unit interval:

$$\hat{s}_{\text{soft}} = \frac{\sum_{k=0}^{5} \hat{p}_k\, k}{5}.$$

Temperature scaling of the soft model Although the OLL objective learns a full distribution, the raw logits are still mis-calibrated (Zhang et al., 2021). We therefore apply slice-wise *temperature scaling* that is **aware of the ordinal distances**. For every slice (u, q) we find $T^{\star}_{(u,q)}$ that minimises distance-aware Earth-Mover distance on the development set:

$$T^{\star}_{(u,q)} = \arg\min_{T > 0} \; \delta - \mathrm{EMD}\big(\mathrm{softmax}\big(\mathbf{z}/T\big), \mathbf{y}_{\text{dev}}\big)$$

where $\mathbf{z}$ are the pre-softmax logits. At inference we divide the logits by the stored $T^{\star}_{(u,q)}$, producing calibrated probabilities without modifying the encoder weights.

For every model: 1. baseline `STS-B RoBERTa-large`, 2. the hard-label variant, and 3. the soft-label variants, we evaluate regression and soft performance on the mode × tercile grid as follows: first, for each (mode, tercile) cell we compute the metric averaged over all items in that cell. Second, within each tercile we macro-average these cell scores across modes, yielding one value per tercile. Finally, we take the mean of the three tercile scores to obtain the overall result reported in Table 1. Additionally, we also report traditional Spearman correlation scores (ρ) on the test portions of Soft-STS-15 and STS-Benchmark. Results are collected in Table 1.

System	δ-EMD ↓				RMSE ↓				Spearman ρ (%) ↑	
	T1	T2	T3	Avg.	T1	T2	T3	Avg.	STS-B	S-STS-15
Baseline	0.0493	0.0280	0.0233	0.0335	0.1364	0.1047	0.1182	0.1197	**91.44**	92.17
Hard	**0.0481**	0.0264	0.0188	0.0311	0.1328	**0.1032**	**0.1073**	0.1144	89.68	94.25
Soft	0.0564	0.0246	0.0216	0.0342	0.1345	0.1041	0.1162	0.1183	88.70	93.77
Soft-Cal	0.0486	**0.0223**	**0.0178**	**0.0296**	**0.1123**	0.1084	0.1219	**0.1142**	89.14	**95.12**

Table 1: Performance on Soft-STS-15 by disagreement tercile. Columns 2–5 report distance-aware calibration error (δ-EMD); columns 6–9 show root mean square error on the mean label; columns 10–11 list rank correlation on STS-B and Soft-STS-15. *Soft-Cal*, i.e. soft-label training plus δ-aware temperature scaling, yields the best overall calibration and the lowest average RMSE, while the hard model remains strongest on the clearer slices (T2–T3 RMSE).

6 Results

We analyze the results presented in Table 1 in relation to our research questions, contrasting distance–aware EMD (calibration) and RMSE (accuracy) across disagreement terciles.

RQ1: Calibration. We first asked whether modeling the full distribution improves distance-aware ordinal calibration (δ-EMD). We observe significant improvements using both proposed strategies: The regression-trained HARD model, enhanced with a simple slice-aware truncated Gaussian head, reduces the calibration error of the STS-B baseline by $\sim$ 7% (0.0311 vs 0.0335). This demonstrates the effectiveness of this computationally light, post-hoc approach which requires no retraining. Furthermore, the *Soft-Cal* model achieves the best global calibration (0.0296), lowering δ-EMD by 12 % relative to the baseline, confirming that the combination of OLL training and δ-aware temperature scaling yields the highest alignment with human judgments, with T2 showing the highest gains (0.0223).

RQ2: Utility Cost. We then examined whether this improved calibration comes at the cost of standard utility metrics. Both *Soft-Cal* and HARD improve significantly over the baseline on Soft-STS-15 Spearman ρ (95.12% and 94.25% respectively, vs 92.17%). Interestingly, this ranking advantage is inverted on the pre-filtered, high-agreement STS-B dataset, with the baseline model obtaining the best result (91.44%), indicating that rankings on curated benchmarks can overestimate a model's robustness by not testing its ability to handle real-world ambiguity.

RQ3: Training Strategies and Trade-offs. Finally, we explore the trade-offs between our two proposed strategies (Soft-Cal vs. Hard). A key finding is the strength of the proposed truncated Gaussian head for reconstructing human label variation in the Soft-STS-15 dataset. First, the HARD model notably outperforms the uncalibrated *Soft* model (0.0311 vs 0.0342 δ-EMD), highlighting that this simple strategy can indeed be more effective than a naive soft-training approach. Second, a clear trade-off emerges. While the Soft-Cal model is better calibrated on clearer data (T2 and T3), the Hard model is more accurate in its point predictions on these same slices (e.g., 0.1073 vs. 0.1219 RMSE on T3). Conversely, on the most ambiguous pairs (T1), the Soft-Cal model is substantially more accurate (0.1123 RMSE), which contributes to its superior overall ranking performance.

7 Conclusion

In this paper, we revisited STS from the perspective of the LeWiDi paradigm, confirming that modeling the full distribution of human annotations significantly improves model calibration. We introduced two effective strategies—a lightweight Truncated Gaussian (TG) head for regression models, and a distance-aware objective (OLL) with temperature scaling—for capturing disagreement in both regression and soft-trained models, and compared them to a strong baseline trained on the high-agreement STS-B dataset. Results show that both approaches outperform this baseline on several metrics. Futhermore, while the raw model trained on the soft labels did not show significant gains over the Hard model, temperature scaling (*Soft-Cal*) did improve calibration to human uncertainty, making it the best performer overall. Despite these promising findings, future work should test if this calibration benefits downstream tasks like NLI or IR.

Limitations

We acknowledge several limitations in our study. First, although the LeWiDi literature offers numerous distribution-aware objectives, we explore only ordinal log-loss and a single post-hoc, δ-aware temperature-scaling scheme; a broader sweep, including other loss functions tested by Castagnos et al. (2022) or Uma et al. (2020) could yield different results. Furthermore, every experiment fine-tunes the same RoBERTa-LARGE cross-encoder; architectural sensitivity to dual encoders or instruction-tuned backbones remains to be verified. Second, our study is conducted solely on the Soft-STS-15 dataset, as it is the only publicly available English STS benchmark with the necessary disaggregated multi-annotator scores. While we prioritize grounding our analysis in authentic human disagreement, the generalizability of our findings is constrained by this single dataset. We strongly encourage the release of more disaggregated datasets to advance research in this area. Finally, our slice-wise calibration method learns specific temperatures for each (mode, tercile) slice. This approach assumes that slice membership—which depends on human disagreement levels—is known at test time. In practical deployments, this metadata is unavailable for new, unlabeled data. However, several practical strategies can address this limitation: (1) **Slice Prediction**: Training a lightweight auxiliary classifier to predict the disagreement level (tercile) of a new sentence pair; (2) **Ensemble-based Estimation**: Using model ensembles to dynamically estimate uncertainty as a proxy for disagreement; or (3) **Global Fallback**: Applying a single, globally-trained temperature optimized across the entire development set.

Acknowledgments

This work is partially funded by the Spanish Ministry of Science, Innovation and Universities (project FairTransNLP PID2021-124361OB-C32) funded by MCIN/AEI/10.13039/501100011033, by a research agreement between UNED and Red.es (C039/21-OT-AD2) and by ERDF, EU A way of making Europe. Alejandro Benito-Santos' work was supported by grant JDC2022-048408-I funded by MICIU/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR". However, the views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics. 116 citations (Crossref) [2023-06-21] tex.ids= agirre_semeval-2015_2015.

Valerio Basile. 2021. It's the End of the Gold Standard as We Know It. In *AIxIA 2020 – Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 441–453, Cham. Springer International Publishing. 2 citations (Crossref) [2023-10-04] tex.ids= basile_its_2020.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A Simple Log-based Loss Function for Ordinal Text Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics. 268 citations (Crossref) [2023-06-21] tex.ids= cer_semeval-2017_2017.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–86, Singapore. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. *Departmental Papers (ASC)*. Tex.ids= krippendorff_computing_2011-1.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 3 citations (Crossref) [2023-10-16].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474, Red Hook, NY, USA. Curran Associates Inc.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Tetsuya Sakai. 2021. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A Case for Soft Loss Functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8:173–177.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty Estimation and Reduction of Pre-trained Models for Text Regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective Human Opinions in Semantic Textual Similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

Xinran Zhang, Maosong Sun, Jiafeng Liu, and Xiaobing Li. 2021. Optimal Embedding Calibration for Symbolic Music Similarity. *arXiv preprint*. ArXiv:2103.07656 [cs].

A Figures

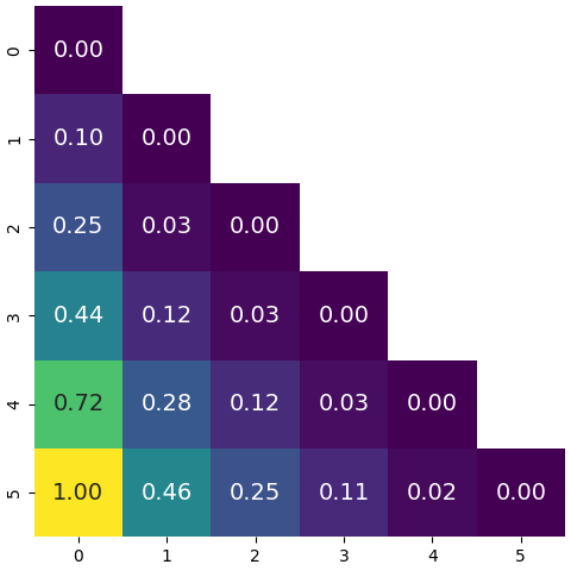

Figure 1: Normalized Krippendorff disagreement matrix $\boldsymbol{D}$ ($D_{jk} \in [0, 1]$). Brighter cells indicate *higher* perceptual distance, i.e. rarer coder confusions for a given pair. Notice the short distances between adjacent pairs (~ 0.03) except for the 0-1 pair (0.10).

B Contingency Table

Mode	Per-item Agreement (α_i) Tercile			Row Σ
	T1	T2	T3	
∅	605 (7.67%)	583 (7.39%)	0 (0.00%)	1 188 (15.06%)
0	712 (9.02%)	357 (4.52%)	1 585 (20.09%)	2 654 (33.64%)
1	583 (7.39%)	353 (4.47%)	162 (2.05%)	1 098 (13.92%)
2	258 (3.27%)	174 (2.21%)	69 (0.87%)	501 (6.35%)
3	309 (3.92%)	458 (5.80%)	88 (1.12%)	855 (10.84%)
4	165 (2.09%)	556 (7.05%)	172 (2.18%)	893 (11.32%)
5	69 (0.87%)	251 (3.18%)	381 (4.83%)	701 (8.88%)
Col Σ	**2 701 (34.23%)**	**2 732 (34.63%)**	**2 457 (31.14%)**	**7 890 (100%)**

Table 2: Contingency table of sentence pairs by majority label (*mode*) and agreement tercile. Counts are accompanied by the percentage of the entire dataset ($N = 7\,890$).