ABDN-NLP at CoMeDi Shared Task: Predicting the Aggregated Human Judgment via Weighted Few-Shot Prompting

Ying Xuan Loke[1] **Dominik Schlechtweg**[2] **Wei Zhao**[1]

[1]University of Aberdeen [2]University of Stuttgart

y.loke.22@abdn.ac.uk

schlecdk@ims.uni-stuttgart.de

wei.zhao@abdn.ac.uk

Abstract

Human annotation is notorious for being subjective and expensive. Recently, Schlechtweg et al. (2025) introduced the CoMeDi shared task aiming to address this issue by predicting human annotations on the semantic proximity between word uses, and estimating the variation of the human annotations. However, distinguishing the proximity between word uses can be challenging, when their semantic difference is subtle. In this work, we focus on predicting the aggregated annotator judgment of semantic proximity by using a large language model fine-tuned on 20 examples with various proximity classes. To distinguish nuanced proximity, we propose a weighted few-shot approach that pays greater attention to the proximity classes identified as important during fine-tuning. We evaluate our approach in the CoMeDi shared task across 7 languages. Our results demonstrate the superiority of our approach over zero-shot and standard few-shot counterparts. While useful, the weighted few-shot should be applied with caution, given that it relies on development sets to compute the importance of proximity classes, and thus may not generalize well to real-world scenarios where the distribution of class importance is different[1].

1 Introduction

Human annotation, which leverages human annotators to create gold-standard labels, has been an essential step when curating training data for machine learning tasks. However, this process is particularly challenging due to the subjective nature of human judgment. Such subjectivity may result in significant disagreements among human annotators, giving rise to poor quality of gold-standard labels—which may further trouble the reliability of models trained on these labels. While many efforts have focused on using aggregation to mitigate disagreements among annotators (Uma et al., 2021; Leonardelli et al., 2023), very few works studied the fundamental aspects of disagreements, such as the complexity and underlying causes that may lead to disagreements in human annotation.

Recently, Schlechtweg et al. (2025) introduced the CoMeDi 2025 shared task, which investigates annotation disagreements in semantic proximity between word uses through two subtasks: (i) predicting the aggregated judgment among human annotators, and (ii) predicting the variation of annotations by estimating the level of disagreement in annotating semantic proximity.

In this work, we focus on the first subtask and build our approaches upon the work by Yadav et al. (2024), which leverages large language models (LLMs) to produce human judgments of semantic proximity. We refer to their approach as automating human judgment. Approaches of this kind have been shown to incur a much lower cost in annotation compared to using human annotators to do so (Gilardi et al., 2023). Our main contribution is to introduce a weighted few-shot learning approach that prompts LLMs to predict human judgments of the proximity class between word uses, on an ordinal scale ranging from 1 to 4, and fine-tunes LLMs on 20 examples to help them learn how such judgments are made. Our few-shot approach differs from the standard one in that important proximity classes receive greater attention during fine-tuning.

2 Task Description

The CoMeDi 2025 shared task explores annotation disagreements through two subtasks, both of which are based on human Word-in-Context (WiC) judgments across seven languages. Each data instance contains a target word w with a pair of uses u_1 and u_2, where each usage conveys a context-specific meaning. Each use pair associates with a human

[1]Our implementation is made publicly available at https://github.com/yingxuaaaaaan/automating-semantic-proximity-annotation

Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 122–128
January 19, 2025.

Target word: chairman
Usage 1: ..out of respect to the chairman's cough...
Usage 2: Ronald J. Gidwitz, chairman, Illinois State Board of Education..

Human judgments: [3, 4, 4]
Median of judgments: 4
Mean pairwise difference of judgments: 0.667

Figure 1: A running example for the target word 'chairman'. The semantic proximity of the two uses are judged by three annotators as context variance (3), identity (4) and identity (4), respectively.

judgment on an ordinal relatedness scale ranging from 1 to 4. The judgment reflects the semantic proximity between a pair of uses, interpreted as homonymy (1), polysemy (2), context variance (3), and identity (4), respectively. An running example is illustrated in Figure 1. The subtask descriptions are outlined as follows:

- Subtask 1: For each use pair (u_1, u_2), participants are asked to predict the median of annotator judgments regarding semantic proximity of the two uses. Predictions are evaluated against the median labels using the ordinal version of Krippendorff's α (Krippendorff, 2018).
- Subtask 2: For each use pair (u_1, u_2), participants are asked to predict the level of annotation disagreement in semantic proximity between the two uses. The level of disagreement is calculated as the mean of pairwise absolute judgment differences among annotators. Predictions are evaluated against the mean disagreement labels using Spearman's ρ (Spearman, 1961).

3 Our System

In this work, our focus is on subtask 1. Our system leverages GPT-4o-mini to predict the aggregated annotator judgment per use pair through prompting. We experiment with three prompting setups: zero-shot, standard few-shot and weighted few-shot.

Zero-shot. Our prompt and model configuration are based on the template by Yadav et al. (2024). The prompt is designed to automate the annotation of semantic proximity by prompting LLMs to follow human annotation guidelines to produce a judgment for each use pair. Additionally, they found that model performance is affected greatly by model hyperparameters such as temperature and top-p, which control the diversity and randomness of the model output. We adopt the model configuration from their work and set both top-p and temperature to 0.9.

Standard few-shot. Our prompt in the standard few-shot setup extends upon the zero-shot prompt by providing a small number of examples for GPT-4o-mini to learn annotator judgments on proximity classes. For instance, in the n-shot setup, we randomly sample n equally sized data instances per judgment (proximity) class from **development data** and incorporate these instances into the prompt. In this case, we assume the four judgment classes are equally important.

Weighted few-shot. Our preliminary results showed that performance gaps between judgment (proximity) classes are substantial (e.g., the judgment class 1 is often the most difficult class for GPT-4o-mini to predict, cf., Figure 5). Additionally, we found that the number of data instances per judgment class is imbalanced (see Figure 4). This indicates that the four judgment classes are not equally important. Based on these observations, we propose a weighted few-shot scheme: we first compute the importance per judgment class, and for each class we randomly sample data instances from **development data** based on the class importance—**the more important a judgment class is, the greater attention it will receive**, i.e. that we will sample many more data instances of that class compared to other classes for fine-tuning GPT-4o-mini. As a result, this approach will prioritize model improvement on important classes. We consider two implementations of class importance, based on: (a) **class frequency** and (b) **class difficulty**. For (a), the importance of each class is estimated based on the percentage of data instances belonging to that class. We use these percentages as probabilities for sampling data instances in each class. Note that we compute importance separately for each language. For (b), we refer the importance of each class to the model performance of that class. To do so, we compute the inverted F_1 score (the harmonic mean of precision and recall) for each class, and normalize it across the four classes, denoted by:

$$p_i = \frac{{F_1}^{-1}(i)}{\sum_{j \in (1,2,3,4)} {F_1}^{-1}(j)}$$

where p_i is the importance of the i-th class that we use as the probability for sampling data instances

belonging to that class from development data. Alternative measures for estimating class difficulty are mostly based on entropy (Capecchi and Moller, 1968; Li et al., 2019; Juszczuk et al., 2021), which we will explore in future work.

Note that we use the raw texts without applying lemmatization or removing punctuation, nor do we explore advanced LLMs such as GPT-4o and Llama 3. Instead, our system focuses on showcasing the use of our weighted few-shot prompting for predicting the aggregated annotator judgment in semantic proximity, and therefore our system performance might be suboptimal.

In the case that prompting GPT-4o-mini does not generate a ordinal judgment class for a use pair, we assign Judgment 0 to that use pair and treat it as an outlier. We note that such cases are very rare in our experiments, and therefore their impact on model performance is expected to be small.

Prompt engineering. Our prompt builds upon the template by Yadav et al. (2024), with the following modifications. Firstly, we provide examples by appending them to the prompt; doing so will not update model weights while Yadav et al. (2024) submit a fine-tuning job to the OpenAI server that will update model weights. Secondly, we restrict the formatting of model response to include the identifiers of each use pair, to which we observe performance gains on development sets. We attribute performance gains to the fact that including identifiers help avoid mismatches between a judgment class prediction and the corresponding use pair. Mismatch may happen in our setup as we prompt GPT-4o-mini in batch, i.e., judgment classes for a batch of use pairs are predicted at once. Note that such identifiers are added to the prompt only in the few-shot setup, as we observe that, without providing examples to fine-tune the model, identifiers are sometimes not generated in model responses.

We additionally experimented with including a language identifier in the prompt to state which language each use pair belongs to, but this is not helpful. Our prompt in the zero-shot setup is displayed in Figure 2. The prompt in the standard and weighted few-shot setups is provided in Figure 3.

[SYSTEM]
You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic similarity of the target word in these sentences 1 and 2. Consider only the objects/ concepts the word forms refer to: ignore any common etymology and metaphorical similarity! Ignore case! Ignore number (cat/Cats = identical meaning). If target is emoji then rate by its contextual function. Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is unrelated; 2 is distantly related; 3 is closely related; 4 is identical in meaning. Your response should align with a human's succinct judgment. Please respond in the format:
[USER]
Keyword (target word): <value>
Sentence 1: <value>
Sentence 2: <value>

Please provide a judgment as a single integer. For example, if your judgment is Identical, then provide 4. If your judgment is Unrelated, provide 1.

Figure 2: Our prompt in the zero-shot setup.

4 Experimental Setup

Datasets. The CoMeDi shared task provides datasets for seven languages: Chinese, English, German, Norwegian, Russian, Spanish, and Swedish. These were sampled from publicly available datasets (Schlechtweg et al., 2018; Schlechtweg, 2023; Schlechtweg et al., 2021; Hätty et al., 2019; Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021; Kurtyigit et al., 2021; Aksenova et al., 2022; Kutuzov et al., 2022; Zamora-Reina et al., 2022; Chen et al., 2023) and supplemented with unpublished data (Schlechtweg et al., 2024). Each dataset is divided into three splits: train, development, and test sets. Table 1 presents the data statistics for these datasets. We observed class imbalance in terms of the percentage of instances per judgment class (see Figure 4).

	Train set		Dev set		Test set	
Languages	#data	#tgts	#data	#tgts	#data	#tgts
Russian	8029	189	1126	28	2285	55
Swedish	5457	30	871	5	1345	9
Spanish	4821	70	621	10	1497	20
Norwegian	4494	56	611	8	1380	16
English	5910	31	863	5	2444	10
Chinese	10833	28	2532	4	3240	8
German	8279	116	1663	17	3141	34

Table 1: Statistics of the CoMeDi datasets. '#tgts' denotes the number of target words; '#data' means the number of use pairs.

Class imbalance. In the zero-shot setup, the imbalance of judgment classes will not harm GPT-4o-mini, as we do not fine-tune the model on the CoMedi datasets. For the standard few-shot setup, we provide equally sized examples to fine-tune the

[SYSTEM]
You are a highly trained text data annotation tool capable of providing subjective responses. Rate the semantic similarity of the target word in these sentences 1 and 2. Consider only the objects/ concepts the word forms refer to: ignore any common etymology and metaphorical similarity! Ignore case! Ignore number (cat/Cats = identical meaning). If target is emoji then rate by its contextual function. Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is unrelated; 2 is distantly related; 3 is closely related; 4 is identical in meaning. Your response should align with a human's succinct judgment. Please respond in the format:

Identifier1: <value>
Identifier2: <value>
Rating: <value>

Examples
[USER]
Identifier1: <value>
Identifier2: <value>
Keyword (target word): <value>
Sentence 1: <value>
Sentence 2: <value>

[ASSISTANT]
Identifier1: <value>
Identifier2: <value>
Rating: <value>

Figure 3: Our prompt in the few-shot setup.

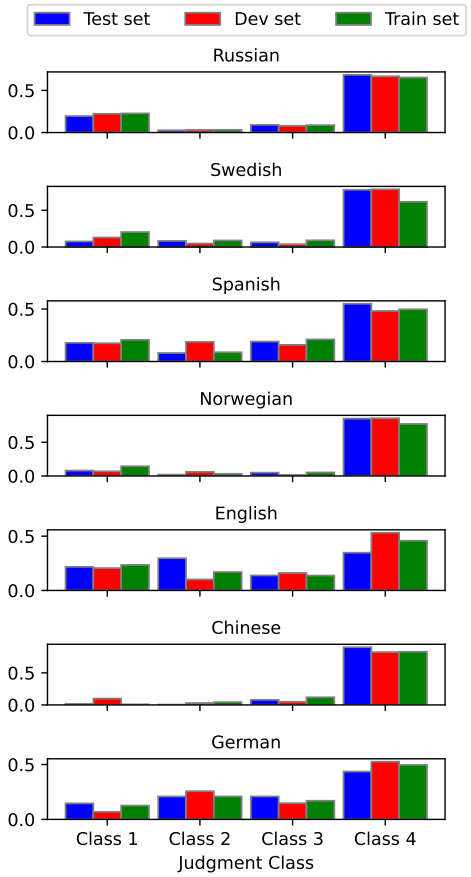

Figure 4: Class frequency distributions across train, dev and test sets, where y-axis shows the percentage of instances per judgment class.

model via in-context learning, aiming to avoid sampling bias stemming from data imbalance.

However, we hypothesize that equally sized sampling is suboptimal because it does not make use of prior knowledge from developments sets, such as class frequency and difficulty distributions. Integrating such knowledge into the few-shot learning process might be useful. For instance, if judgment class 4 is the most popular or most difficult class, providing more examples of that class to fine-tune the model would prioritize model improvement on important classes. Nevertheless, there is no guarantee that the class frequency and difficulty distributions are the same (or comparable) across data splits, but we assume that the difficulty distribution is more consistent than the frequency distribution across splits, as the test set could contain any number of instances per judgment class while the class difficulty reflects its inherent complexity, less affected by data splits.

Results. Table 2 compares our approach in various setups on the CoMeDi test set for the post-evaluation subtask 1. Overall, our approach based on GPT-4o-mini in the zero-shot setup yields moderate Krippendorff scores in most cases, indicating moderate agreement between model and human judgments in semantic proximity. We see our approach performs poorly in Norwegian and Chinese, meaning that GPT-4o-mini may struggle to understand these two languages.

Secondly, we see "standard few-shot", which fine-tunes GPT-4o-mini on totaling 20 examples across 4 classes through in-context learning, is useful. It outperforms the counterpart in the zero-shot setup on average (0.403 vs. 0.388). This is not surprising, as few-shot learning help GPT-4o-mini learn how human judgments are made. Additionally, we observe that our weighted few-shot approach relying on 'frequency' achieves the best performance on average among the four setups. This is because class frequency distributions are generally consistent in both dev and test sets (see Figure 4). In contrast, we see the weighted few-shot relying on 'difficulty' performs only slightly better than 'standard few-shot', which we attribute to the fact that class difficulty distributions differ

Setup	Russian	Swedish	Spanish	Norwegian	English	German	Chinese	Avg
zero-shot (n=0)	0.504	0.351	0.491	0.207	0.610	0.529	0.026	0.388
standard few-shot (n=20)	0.423	0.441	**0.587**	0.197	**0.626**	0.675	-0.127	0.403
weighted few-shot (frequency, n=20)	0.478	**0.509**	0.569	**0.431**	0.625	0.673	**0.209**	**0.499**
weighted few-shot (difficulty, n=20)	**0.512**	0.389	0.543	0.183	0.600	**0.690**	-0.056	0.408
deep-change (Kuklin and Arefyev, 2025)	**0.623**	**0.675**	**0.748**	**0.668**	**0.732**	**0.723**	**0.424**	**0.656**
comedi-baseline (Schlechtweg et al., 2025)	0.112	0.018	0.175	0.124	0.102	0.274	0.059	0.123

Table 2: Krippendorff's results from GPT-4o-mini on the test set in the post-evaluation CoMeDi subtask 1. "deep-change" is the best-performing system in the CoMeDi leaderboard.

across data splits to a large degree (see Figure 5).

Our approach, even in the zero-shot, performs much better than comedi-baseline—which relies on XLM-R coupled with a threshold-based classifier tuned on training data. This means prompting LLMs could yield very competitive results. However, our approach lags behind deep-change—which fine-tunes the Word-in-Context model on the training data of the shared task; this is because deep-change benefits greatly from fine-tuning on the full training data that is 300-500 times larger than the number of training examples we provided in the few-shot setups.

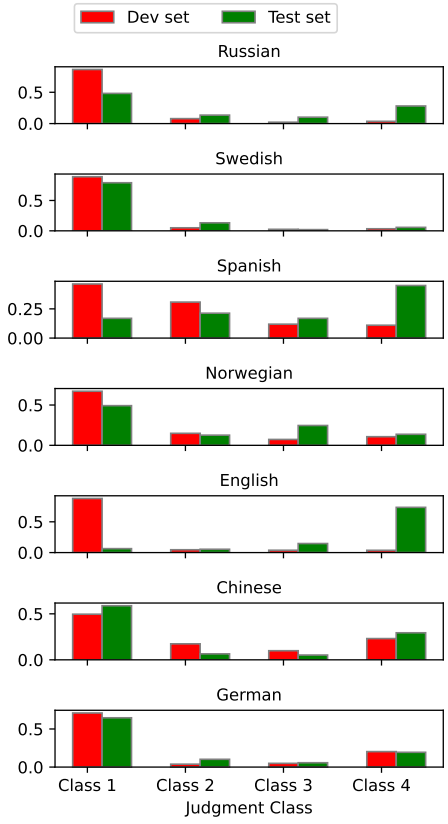

Figure 5: Class difficulty distributions across train, dev and test sets, where y-axis shows the inverted F_1 score per class after normalization.

5 Conclusions

In this work, we leverage a large language model to predict the aggregated human judgment of the semantic proximity between word uses. In particular, we explore several few-shot learning approaches for the model to learn annotator judgments through fine-tuning. Our results demonstrate that our weighted few-shot approach outperforms standard few-shot and zero-shot approaches.

Limitations. In the shared task setup, the class frequency distributions generally are consistent across data splits for all languages. However, such alignment is not guaranteed in real-world scenarios. If distributions differ across splits, performance gains from weighted few-shot learning may become small or even disappear. While class difficulty distributions might be consistent and are not affected much by data splits, but giving greater attention to difficult classes may not be useful in the case that such classes are rare in test sets. As such, how best to leverage prior knowledge (class difficulty and frequency distributions) does not have a straightforward answer, and the standard few-shot learning is still useful when the reliability of prior knowledge is uncertain. Additionally, our findings are based on a single LLM and might differ when we use other LLMs. Moreover, our approach is suboptimal: further improvements could benefit from cleaning up datasets, using stronger LLMs, fine-tuning on a large number of examples in few-shot setups, developing a new approach combining both class frequency and difficulty factors, and others.

Acknowledgements

We thank the anonymous reviewers for their thoughtful feedback that greatly improved the texts. Dominik Schlechtweg has been funded by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Anna Aksenova, Ekaterina Gavrishina, Elisei Rykov, and Andrey Kutuzov. 2022. RuDSI: Graph-based word sense induction dataset for Russian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 77–88, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Vittorio Capecchi and Frank Moller. 1968. Some applications of the entropy to the problems of classification. *Quality & Quantity*, 2.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change*, Singapore. Association for Computational Linguistics.

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.

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA.

Przemysław Juszczuk, Jan Kozak, Grzegorz Dziczkowski, Szymon Głowania, Tomasz Jach, and Barbara Probierz. 2021. Real-world data difficulty estimation with the use of entropy. *Entropy*, 23(12):1621.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Mikhail Kuklin and Nikolay Arefyev. 2025. Deepchange at CoMeDi: the cross-entropy loss is not all you need. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical Semantic Change Discovery. 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)*, Online. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Lusi Li, Haibo He, and Jie Li. 2019. Entropy-based sampling approaches for multi-class imbalanced problems. *IEEE Transactions on Knowledge and Data Engineering*, 32(11):2159–2170.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. The CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles Spearman. 1961. The proof and measurement of association between two things.

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.

Sachin Yadav, Tejaswi Choppa, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4. *arXiv preprint arXiv:2407.04130*.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.