LLM-Based Zero-Shot Soft Labeling for Anticipating Disagreement in Negotiation Dialogues

Ken Watanabe[♠] **and Katsuhide Fujita**[♣]
[♠]Graduate School of Engineering, Tokyo University of Agriculture and Technology
[♣]Institute of Engineering, Tokyo University of Agriculture and Technology
[♠]watanabe@katfuji.lab.tuat.ac.jp
[♣]katfuji@cc.tuat.ac.jp

Abstract

Negotiation involves complex emotional and strategic dynamics that pose challenges for AI agents in negotiation dialogues. This paper proposes ZeSTEM, a zero-shot soft-labeling method using a large language model-based embedding model, to represent negotiation utterances as graded distributions over possible dialogue acts. This formulation is motivated by the fact that negotiation utterances often realize indirect, mixed, and context-dependent communicative functions that are difficult to capture with rigid one-hot labels. Furthermore, it examines the performance of predictive model training on rule-based annotated hard and soft labels obtained by the proposed method for the task of predicting whether agreement will be reached from partial dialogues, namely, final disagreement anticipation in negotiation mid-dialogues (FDANMD). Soft labeling obtained by the proposed method showed a maximum HIT@3 score of 0.87 against rule-based annotated hard labels, whereas failure cases also demonstrated the limitations of rule-based annotation. Furthermore, using ROC AUC, evaluations of FDANMD across three datasets (CB, DN, and JI) with negotiation progress rates of 0.25, 0.5, and 1.0 revealed that soft labeling is particularly effective at low negotiation progress rates and also offers superior performance on individual datasets and unseen datasets for models trained on multiple datasets. These results motivate the use of soft labeling to incorporate the complexity of negotiation dialogues into intermediate representations and support the generalizability of zero-shot soft labeling and generalizable predictors across a wide range of negotiations beyond known domains.

1 Introduction

Negotiation is an indispensable activity in social life, where each party naturally seeks to maximize its own benefits. However, the inherent complexity of negotiation problems—compounded by negotiators' emotions and diverse social backgrounds often makes it difficult for both sides to reach a mutually satisfactory conclusion. In response to these challenges, researchers have begun to develop negotiation dialogue systems that can either negotiate autonomously on behalf of humans or provide support to human negotiators during the process.

In recent years, the advent of large language models has driven significant progress across numerous tasks in the field of natural language processing (NLP). This progress has also spurred research into the use of large language models (LLMs) for negotiation dialogue systems. Despite these advances, studies (Schneider et al., 2023; Bianchi et al., 2024) have revealed that LLMs possess inherent limitations when it comes to negotiating effectively. Moreover, methods (Deng et al., 2023) that incorporate supervised-learning-based instruction mechanisms—designed for various task-oriented dialogue systems, including negotiation—tend to be less effective for negotiation tasks than for other applications. This discrepancy is likely attributable to the intrinsic difficulties of negotiation, underscoring the need for approaches that are specifically tailored to its unique challenges. For instance, automated negotiation agents and LLMs equipped with instruction mechanisms inspired by game theory concede more than necessary during negotiations, thereby leading to agreements that are disadvantageous.

To address these issues, we aim to preprocess negotiation dialogues so that LLM-based automated negotiation dialogue agents can accurately grasp the situation. Here, we propose a zero-shot soft-labeling method using LLM-based text embedding models that targets dialogue acts expressing the intentions implied in each sentence and does not require prior dataset collection or annotation. In addition to comparing the soft-labeling results with existing annotated datasets, we also compare

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1033–1051

its performance on Final Disagreement Anticipation in Negotiation Mid-dialogues (FDANMD). FDANMD predicts whether a negotiation will ultimately end in disagreement by analyzing dialogues exchanged during the negotiation process. Such anticipation is useful not because an agent should mechanically force agreement or disagreement, but because it can provide an intermediate signal for negotiation support. For example, when the predicted risk of final disagreement increases in the middle of a negotiation, an assistant can recommend clarification, issue reframing, concession timing, or confirmation of latent constraints. In autonomous negotiation, the same signal can be used as a diagnostic representation for deciding whether to ask follow-up questions, avoid premature rejection, or revise a proposal strategy. Comparing performance on this prediction task also allows us to examine whether the key elements necessary for reaching agreement are accurately represented. Furthermore, because this zero-shot preprocessing method does not require prior dataset collection or annotation, we aim to extend the applicability of FDANMD beyond preannotated datasets.

The objectives of this research are as follows:

- Verify the accuracy of the zero-shot soft-labeling method using LLM-based text embedding models.
- Compare the performance of FDANMD on existing hard labels and our soft labels.

The structure of this paper is as follows. Section 2 introduces related work. Section 3 explains the method proposed in this study. Section 4 explains the experiments conducted in this study. Section 5 presents and discusses the experimental results. Section 6 concludes the paper and outlines future research directions.

2 Related Works

2.1 Soft Labeling in NLP

Recent studies in NLP have increasingly questioned the practice of collapsing inherently ambiguous annotations into a single hard label. In tasks where annotator disagreement reflects genuine ambiguity or perspectival variation rather than annotation noise, preserving label distributions or uncertainty signals has been shown to improve modeling fidelity and downstream performance.

For example, Meissner et al. (2021) showed in natural language inference that training on annotator label distributions, rather than gold hard labels, yields better representations and improves the modeling of ambiguity.

Similarly, Fornaciari et al. (2021) demonstrated that soft-label multitask learning can leverage annotator disagreement as useful supervision and consistently outperform corresponding single-task baselines across several NLP settings. Wu et al. (2023) further reported that even for single-label classification, classifiers trained with soft labels can outperform hard-label baselines and exhibit better calibration.

2.2 Negotiation Dialogue Datasets

For modeling strategic dialogues, several corpora of negotiation dialogues are available, such as CRAIGSLISTBARGAIN (CB), a corpus for negotiating product prices, proposed by He et al. (2018); DEALORNODEAL (DN), a corpus for allocating items, proposed by Lewis et al. (2017); and JOBINTERVIEW (JI), a corpus for multi-issue job interviews, proposed by Yamaguchi et al. (2021).

In addition to these, there are CASINO, a corpus for allocating items similar to DN, proposed by Chawla et al. (2021); NEGOCOACH, a corpus on negotiating product prices similar to CB, proposed by Zhou et al. (2019); and PERSUASIONFORGOOD, a corpus on persuasion for donations, proposed by Wang et al. (2019).

2.3 Final Disagreement Anticipation in Negotiation Mid-Dialogues

A dialogue-act-based method for disagreement detection in negotiation dialogues (DDND) was proposed by Yamaguchi et al. (2021). Instead of relying solely on text, this approach uses the sequence of dialogue acts following a specified flow (dialogue act flow) within utterances. The study compared nine models, including:

- Logistic Regression: using bag-of-words with TF-IDF for text and dialogue acts (LR-BOW$_{\text{TEXT}}$ and LR-BOW$_{\text{TAG}}$)
- Gated Recurrent Unit (GRU (Cho et al., 2014))-based models: with a simple linear layer and an added self-attention mechanism for both text and dialogue acts (GRU$_{\text{TEXT}}$, GRU-Att$_{\text{TEXT}}$, GRU$_{\text{TAG}}$, and GRU-Att$_{\text{TAG}}$)

- BERT (Devlin et al., 2019)-based models: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ for text inputs only
- Dummy model: a baseline prediction based on class distribution (Random)

Results showed that for the CB and DN datasets, $\text{BERT}_{\text{BASE}}$ achieved the highest average precision, with $\text{BERT}_{\text{LARGE}}$, GRU_{TEXT}, $\text{GRU-Att}_{\text{TEXT}}$, and GRU_{TAG} performing similarly. However, on the JI dataset, dialogue-act-based models (GRU_{TAG} and $\text{GRU-Att}_{\text{TAG}}$) significantly outperformed the others.

Watanabe and Fujita (2025) proposed Final Disagreement Anticipation in Negotiation Mid-Dialogues (FDANMD), which extended DDND, and applied fine-tuning to BERT-based models for text inputs and GRU-based models for dialogue-act inputs. Experimental results indicate that the fine-tuning method affects training time when the negotiation progress rate is 1.0, which is the same setting as DDND, and affects training time and anticipation performance when the negotiation progress rate is 0.25. They also showed that GRU-based models for dialogue-act inputs perform better than BERT-based models for text inputs when the negotiation progress rate is 0.25, especially on the JI dataset, suggesting that dialogue acts help models better understand the indicators of disagreement.

A similar task, the Dialogue Breakdown Detection Challenge (DBDC), proposed by Higashinaka et al. (2016), aims to improve the consistency of a dialogue system and evaluates whether the system's response is valid when a dialogue history between a human and the system is given. Another similar task, proposed by Heddaya et al. (2023), aims to anticipate which side the outcome of bilateral negotiations would favor. Therefore, these tasks differ from FDANMD, which focuses on whether negotiations will end in disagreement or agreement.

3 Zero-Shot Soft-Labeling Method Using LLM-Based Text Embedding Models

3.1 Overview

Human annotation is time-consuming and financially expensive, making preannotation difficult, especially for large-scale datasets. Furthermore, rule-based annotation makes it difficult to design rules that cover a wide variety of utterances and classify them accurately. Annotation using pretrained models has also been proposed in recent years, but although it is less costly than human annotation, it is still costly, and it is unclear whether it can handle negotiation dialogues outside the training data, such as those in different negotiation domains.

In this study, we propose a zero-shot soft-labeling method using LLM-based text embedding models (ZeSTEM) to assign soft labels to each sentence in negotiation dialogues. This method is expected to be low-cost and cover a wide range of representations and domains because of the expressive power inherent in LLM-based text embedding models.

In this paper, a hard label refers to a discrete dialogue-act annotation, typically represented as a one-hot vector or a small set of rule-triggered categorical labels. In contrast, a soft label refers to a continuous-valued vector over the dialogue-act inventory, where each dimension indicates the degree to which an utterance is associated with a possible act. Thus, the proposed soft labels are not supervised probability distributions obtained from multiple annotators, but embedding-based similarity distributions that preserve graded associations between an utterance and multiple possible acts.

We also propose FDANMD using soft labels as an application example of ZeSTEM. Previous research has achieved FDANMD by training a machine learning model using annotated hard labels, but by using soft labels obtained by ZeSTEM, we hope to realize a versatile FDANMD that goes beyond the scope of annotated datasets.

Dialogue-act labels provide a practical abstraction of utterance functions in dialogue, but negotiation utterances often involve a gap between surface form and communicative force. From the perspective of speech act theory (Austin, 1962; Searle, 1975), an utterance can carry an illocutionary force that differs from its literal form:

- a question may function as a proposal.
- a clarification request may signal resistance.
- an apparent agreement may merely acknowledge a breakdown.

Such indirect speech acts and mixed communicative functions are especially common in negotiation, where speakers strategically manage offers, refusals, concessions, and relational alignment. This motivates representing each utterance not as a single rigid act, but as a graded distribution over possible acts.

3.2 Details of ZeSTEM

There have been many attempts to annotate negotiation dialogues and perform NLP tasks based on abstract information, but as mentioned above, various methods and difficulties are involved in obtaining this abstract information. This is due to the ambiguity inherent in natural language and the wide range of content across negotiation domains. Therefore, ZeSTEM proposes the use of abstract information that preserves this ambiguity. Soft labeling of utterances in negotiation dialogue using ZeSTEM is performed in the following flow:

1. First, we vectorize the text (words or sentences) representing the hard labels using a text embedding model.

2. Next, each utterance in the dialogue is vectorized together with the instruction using an embedding model.

3. Then, we calculate the cosine similarity between the embedding vectors of the obtained hard-label representations and the embedding vectors of the utterances. The cosine similarity corresponding to each hard label is stored as a soft label for that utterance as a vector with the same dimensionality as the number of hard labels.

Although the computational procedure is intentionally simple, the main contribution of ZeSTEM lies in using label-text embeddings as a zero-shot interface between dialogue-act inventories and negotiation utterances. This design enables the same pipeline to be applied to new negotiation domains without retraining an utterance classifier or collecting task-specific annotations. Moreover, by retaining the full similarity vector rather than selecting only the top label, ZeSTEM preserves indirect and mixed communicative functions that are often discarded by rule-based hard annotation.

3.2.1 Vectorizing Hard Labels

The vectors corresponding to the hard labels are used repeatedly afterward, so they are vectorized and stored before the dialogue is processed. The text representing a hard label when vectorized can be the hard-label word itself, a simple sentence such as "This sentence is (label)," or a sentence with a definition or explanation such as "This sentence is (label) with the following properties: ".

3.2.2 Vectorizing Utterances

The dialogue is divided into utterances, and each utterance is vectorized. The target utterance is input to the embedding model during vectorization, and it can also be input together with the previous utterance to clarify the context. The input instructions can range from simply presenting the label category (emotion, strategy, etc.) to explicitly instructing classification by clearly stating the target label and adding definitions and explanations for the target label.

3.2.3 Calculating the Cosine Similarity and Storing It as a Soft Label

Cosine similarity, which represents the similarity between vectors, is used as a score to represent the relevance between vectorized utterances and hard labels. Cosine similarity takes values from -1 to 1, with values closer to 1 indicating higher similarity and values closer to -1 indicating lower similarity. To preserve cosine similarity, all values in the soft label fall within the range from -1 to 1.

3.3 Details of FDANMD with Soft Labels

As mentioned in Section 2.3, a prior study by Watanabe and Fujita (2025) has shown that performance in anticipating final disagreement in the early stages of negotiations was better with a GRU trained on rule-based annotations than with BERT trained on dialogue text. This is thought to be because it was difficult for BERT trained on text to obtain appropriate abstract information from dialogue. However, hard labels substantially reduce information from the original dialogue.

Therefore, we propose FDANMD using soft labels, which are expected to preserve more information than the hard labels presented in prior studies and therefore store more appropriate abstract information.The soft labels arranged in the order of utterances in the dialogue are used as multivariate time-series data to solve the classification problem.

4 Setup

4.1 Evaluating the Performance of ZeSTEM

4.1.1 Datasets

As datasets for evaluating ZeSTEM, we used three types of dialogue-act hard-labeled datasets—CB, DN, and JI from those described in Section 2.2. Table 1 shows a quantitative comparison of the three negotiation datasets: CB, DN, and JI.

Table 1: Quantitative comparison of the three negotiation datasets. In negotiation roles, Sym means that both negotiators have symmetrical roles, and Asym means that both negotiators have asymmetrical roles in the negotiation. ⋆ denotes that some of the issues are interdependent. In this case, two of the five issues in JI are interdependent.

	CB	**DN**	**JI**
# of dialogues	5987	6251	2577
Avg. turns	7.53	4.97	12.7
Agreed [%]	74.9	76.2	92.9
Number of issues	1	3	5⋆
Negotiator roles	Asym	Sym	Asym

Each dataset is annotated with the following dialogue acts:

- <greet> (includes *hi*, *how are you*, etc.)
- <inquire> (includes *what*, *where*, etc.)
- <inform> (reply to <inquire>)
- <propose> (includes *I'd like*, etc.)
- <agree> (includes *deal*, *that works*, etc.)
- <disagree> (includes *can't*, *worse*, etc.)

These are annotated using regular expressions, and utterances that do not fit any of these categories are annotated as <unk> (unknown).

4.1.2 Models

The text embedding models used in ZeSTEM are Qwen3 Embedding 0.6B and Qwen3 Embedding 8B (Zhang et al., 2025). Both were adopted because they are general-purpose embedding models that embed instruction sentences and target text.

4.1.3 Variations of Texts to Be Vectorized

First, we compared six patterns of texts to be vectorized in total: three methods for converting label information into text for input into the model—the label word itself, a simple sentence, or a detailed sentence with an explanation—and two patterns for inputting the utterance to be soft-labeled into the embedding model: either inputting only the utterance to be soft-labeled or inputting it together with the preceding utterance. Instructions input along with the utterance to be soft-labeled do not include descriptions of the labels.

Furthermore, based on the results, we compared three patterns of instructions to be input along with the utterance to be soft-labeled: instructions that do not include descriptions of the labels, instructions that list the label names, and instructions that list the label names and their descriptions.

A complete list of transcribed labels and instructions can be found in Appendix A.

Table 2: An example of a negotiation dialogue. Turn indicates the turn number. r indicates the negotiation progress rate. Utterance indicates the content of the utterance.

Turn	r	Utterance
1	0.25	I want the red one. How about $5?
2	0.50	I need at least $7.
3	0.75	Hmm... How about $6.5?
4	1.00	OK. Deal.

4.1.4 Metrics

We evaluate by the degree of match with rule-based annotated hard labels. Since a single utterance can be hard-labeled with multiple hard labels (up to three), we evaluate using HIT@3, which assesses whether each hard label is ranked among the top similarities, and HIT@1, which assesses whether the most similar hard label matches that label when there is only one annotation, or whether it matches one of the hard labels when there are multiple annotations.

4.2 Evaluating the Performance of FDANMD Using Soft Labels by ZeSTEM

4.2.1 Tasks

Table 2 shows an example of a negotiation dialogue. Based on a dialogue containing n turns of utterances between two negotiators, the model determines whether an agreement has been reached (successful) or not (failed).

The negotiation dialogue D comprises $n \in \mathbb{N}$ turn utterances $\{s_1, s_2, \ldots, s_n\}$, and an utterance s_i ($i \in [1, n]$) is made by one negotiator and contains one or more sentences. Let D' be the mid-dialogue consisting of the opening $[rn]$ turns of D. Considering D' as the input, whether the negotiation dialogue D succeeds or fails is determined; success is labeled 0 and failure is labeled 1.

In this study, the negotiation progress rate r was set to 1.0, 0.5, and 0.25. This follows prior research shown in Section 2.3. These values correspond to the end, middle, and early stages of negotiations, respectively.

4.2.2 Datasets

The three datasets with hard-labeled dialogue acts (CB, DN, and JI) are also annotated with final agreements and disagreements and were used in prior studies shown in Section 2.3, so we use these as the ground truth. To assess generalizability, we examine predictive performance on unseen datasets that were not used for training and on individual datasets when trained on mixtures of multiple datasets. Therefore, there are seven variations of the training dataset: CB only, DN only, JI only, CB and DN, CB and JI, DN and JI, and CB, DN, and JI, respectively, and we evaluated each on all three datasets: CB, DN, and JI.

4.2.3 Models

The MiniRocket transformer (Dempster et al., 2020, 2021) was used to transform the input sequence, and the Ridge Classifier (Hoerl and Kennard, 1970) was used as the classification model for FDANMD. By comparing the performance of lightweight classifiers, we clearly confirm the applicability of soft labels as abstract information.

4.2.4 Metrics and Baselines

Due to the large class imbalance in whether agreement or disagreement was reached, accuracy is high—especially at 92.9% in JI—even if a model predicts that all dialogues end in agreement. Therefore, as in prior research, this study also evaluates using ROC AUC (Bradley, 1997).

We performed five-fold cross-validation throughout the experiment to reduce the instability of the results due to insufficient data.

For comparison, we use annotated hard labels vectorized using one-hot encoding. By standardizing the input format to the model, we minimize variations other than the input information.

4.3 Experimental Environment

4.3.1 Hardware

One Nvidia RTX 5090 was used to run Qwen3 Embedding. Since the model parameter file size is 16GB even for the 8B model, experiments at BF16 precision are possible with a single RTX 5090, which has 32GB of VRAM. In addition, other hardware specifications include an AMD Ryzen Threadripper PRO 5995WX 64-Cores CPU and 512GB of RAM.

4.3.2 Software

All programs used in the experiments were written in Python. The Python version was 3.13.12. vLLM (Kwon et al., 2023) was used to implement Qwen3 Embedding. The vLLM version was 0.15.1. sktime (Löning et al., 2019; Király et al., 2025) was used to implement the MiniRocket transformer, and scikit-learn (Pedregosa et al., 2011) was used to implement the Ridge classifier. The sktime version was 0.40.1 and the scikit-learn version was 1.7.2.

5 Experimental Results and Discussion

5.1 Evaluating the Performance of ZeSTEM

Results of ZeSTEM are shown in Table 3. Overall, the 0.6B model consistently exhibits low accuracy in HIT@3 and HIT@1, while some variations of the 8B model show high accuracy, approximately 0.85 in HIT@3. Although HIT@1 reaches only approximately 0.65 even with the 8B model, it is still substantially higher than the 0.6B models, for which all variations fell below 0.5.

Because the target sentences are short, as they are utterances, we initially expected the 0.6B model to be sufficient, but the results were contrary to our expectations. This is likely because the 0.6B model lacked sufficient knowledge about the information contained in utterances within negotiation dialogues and therefore could not acquire appropriate embedding representations.

When comparing methods for converting labels to text, vectorizing the label as just a word or a simple sentence tends to be more accurate than converting it into a sentence along with a detailed explanation of the label.

This result is likely due to the fact that adding detailed explanations to the labels actually caused the embedded representations to deviate from the essence of the labels, widening the gap between them and the actual annotated text.

When comparing inputting only the target utterance versus inputting it together with the preceding utterance, the combination with the preceding utterance appears to be better.

Regarding these results, the interpretation of which dialogue act a given utterance corresponds to is influenced by the preceding utterance. For example, one of the definitions for `<inform>` in rule-based annotation was that the preceding utterance was `<inquire>`. Therefore, it is a natural consequence that inputting the target utterance along with the preceding utterance improves accuracy.

Table 3: Results of evaluating the performance of ZeSTEM for different input forms of utterances and labels. Target Utterance Only and With Preceding Utterance indicate that only the target utterance to be soft-labeled is input into the embedding model and that the target utterance is input together with the preceding utterance, respectively. Word indicates that label information was converted into label words, Simple indicates that label information was converted into simple sentences, and Detailed indicates that label information was converted into sentences with detailed descriptions. Bold text with an underlined line indicates the best possible score for that metric. Underlined text alone indicates the second-best possible score for that metric.

		Target Utterance Only			With Preceding Utterances		
Metrics	Models	Word	Simple	Detailed	Word	Simple	Detailed
HIT@3	0.6B	0.4904	0.6183	0.4572	0.6202	**0.6430**	0.4464
	8B	0.8521	0.8479	0.7195	**0.8594**	0.8538	0.7305
HIT@1	0.6B	0.3648	0.4650	0.1889	0.3892	**0.4899**	0.2104
	8B	0.6270	0.6678	0.5517	0.6500	**0.6692**	0.5065

Table 4: Additional results of evaluating the performance of ZeSTEM with different instructions. Genre indicates that instructions include only the genre to be labeled, Word indicates that instructions include the list of label words, and Detailed indicates that instructions include the list of labels with detailed descriptions. In all of these results, the 8B model was used, with labels embedded as words only and utterances embedded together with the preceding utterance. Bold text indicates the best possible score for that metric.

Metrics	Genre	Word	Detailed
HIT@3	0.8594	0.8441	**0.8707**
HIT@1	**0.6500**	0.6300	0.6402

As an additional experiment based on these results, Table 4 shows the results of investigating how accuracy changes depending on the type of instruction embedded simultaneously with the utterance. In this comparison, the model used is Qwen3 Embedding 8B, with labels embedded as words only and utterances embedded together with the preceding utterance. These results show that, unlike label encoding, for HIT@3, adding the list of labels with detailed explanations to the instructions resulted in higher accuracy. For HIT@1, providing only the genre to be labeled yielded the highest result, but the list with detailed explanations was more effective than the list with only label words.

Regarding these results, it is thought that the detailed explanations in the instructions may have encouraged the embedded representation of the target utterance to become closer to the label.

Based on these results, we used the soft labels generated by vectorizing the labels as words only, the target utterances together with preceding utterances, and the instructions with the list of labels with detailed explanations in the next experiment shown in Section 5.2.

Furthermore, examining the failures in HIT@1 reveals the limitations of rule-based hard labeling. For example, the utterance "sounds pretty nice. would you take $705 for it?" in CB and "well now. i see two of everything. how about we split it down the middle?" in DN both receive the highest score for <propose> in soft labels by ZeSTEM, while the rule-based annotation labels them as <inquire> because they contain "would you" and "how about." In terms of illocutionary force, these utterances are indirect speech acts: although their surface form is interrogative, their communicative function is to make an offer or counter-offer. This explains why a rigid rule based on interrogative expressions maps them to <inquire>, whereas ZeSTEM assigns a higher score to <propose>. Additionally, the DN includes utterances such as "okey it's no deal" and "alright no deal," where agreement is not reached. In contrast, soft labels by ZeSTEM gives the highest score for <agree>, while the rule-based annotation labels them as <disagree> because they contain "no." Regarding this, while <disagree> might seem appropriate because there is disagreement on the content of the negotiation, <agree> is considered appropriate because, based on its relationship with the preceding sentence, there is agreement to conclude that there is disagreement. As shown here, there are differences in dialogue acts due to the content and context of the utterances themselves that cannot be picked up by rule-based annotation, and it can be said that ZeSTEM is able to capture these differences.

Conversely, a tendency for disagreement not to be captured was observed, mainly in JI. For example, regarding the location of work, in response to the suggestion "would it be possible to work

in seoul or syney?", the response "sydney?" is given. While rule-based annotation labels this as <disagree> because the preceding suggestion was not accepted, the soft label by ZeSTEM shows <inquire> as the highest value. The response "sydney?" is formally a short clarification question, but in context it may carry the illocutionary force of resistance or rejection. ZeSTEM interpreted the surface-level inquiry function more strongly than the implicit refusal function. Thus, in some cases, rule-based methods were able to more accurately annotate dialogue acts using expressions that ZeSTEM could not capture.

5.2 Evaluating the Performance of FDANMD Using Soft Labels by ZeSTEM

Results of evaluating the performance of FDANMD using soft labels by ZeSTEM are shown in Figure 1. Detailed results are provided in Appendix B. Comparisons with prior study methods and cutting-edge text-based LLMs are also provided in Appendix C. Overall, anticipation using soft labels performed better than anticipation using hard labels. In particular, in the CB dataset evaluation when the negotiation progress rate was $r = 1.0$, there was a statistically significant difference of $p < 0.01$ for all combinations of training data.

Furthermore, predictions using soft labels tend to perform better on unseen data. When a model trained on one dataset is used to predict another dataset, such as predicting CB with a model trained on DN, performance is not as good as when predicting the untrained portion of the same dataset used for training. However, when a model trained on two datasets is used to predict another dataset, such as predicting CB with a model trained on DN and JI, performance tends to be comparable or even better. This tendency is also seen to some extent in predictions using hard labels, but when $r = 0.25$, it is particularly noticeable with soft labels.

Furthermore, there is a tendency for the scores on each original dataset to be less likely to drop when multiple datasets are mixed. The smaller the negotiation progress rate, the lower the performance of predictions using hard labels when data other than the prediction dataset are mixed into the training dataset. However, it can be confirmed that performance does not significantly drop when training data are mixed in predictions using soft labels, even when the negotiation progress rate is $r = 0.25$.

Particularly, when evaluating with JI, models trained on CB and DN, where JI is unseen data to the model, and models trained on CB, DN, and JI consistently outperform models trained on the JI dataset alone when using soft labels.

These results suggest that representing possible dialogue acts using soft labels was more advantageous for understanding complex negotiation dialogue than fixing dialogue acts using hard labels. While it was difficult for a lightweight model to learn commonalities in the flow of dialogue acts across negotiation domains using predictions from hard labels, predictions from soft labels made it possible for a lightweight model to learn such commonalities. Therefore, evaluation on individual and unseen data by a model trained with multiple datasets varies little from the evaluation of a model trained only with the target data. In particular, because JI has a small number of data items and few disagreements, training with data from multiple different domains appears to have improved performance.

These results should be interpreted in two complementary ways. First, ZeSTEM may correct some surface-pattern errors in rule-based labels by assigning high scores to communicative functions that are implied by context. Second, even when the top-ranked act is not fully correct, the dense similarity vector may provide smoother features than sparse one-hot labels. Therefore, the gains in FDANMD likely arise from both speech-act-sensitive representation and feature smoothing. Disentangling these two factors remains an important direction for future work.

6 Conclusion

We proposed ZeSTEM, a zero-shot soft-labeling method using text embedding models, as an alternative to hard-label annotation.

Testing the accuracy of ZeSTEM with annotated datasets shows that the 8B model performs substantially better than the 0.6B model. Furthermore, the accuracy of the model changes depending on the format of the labels, target utterances, and instruction sentences input into the model.

Performance improved when running the FDANMD task, which previously used hard labels, with soft labels. Furthermore, the soft-label approach showed superior performance on individual datasets and on unseen datasets for models trained on multiple datasets. The performance difference

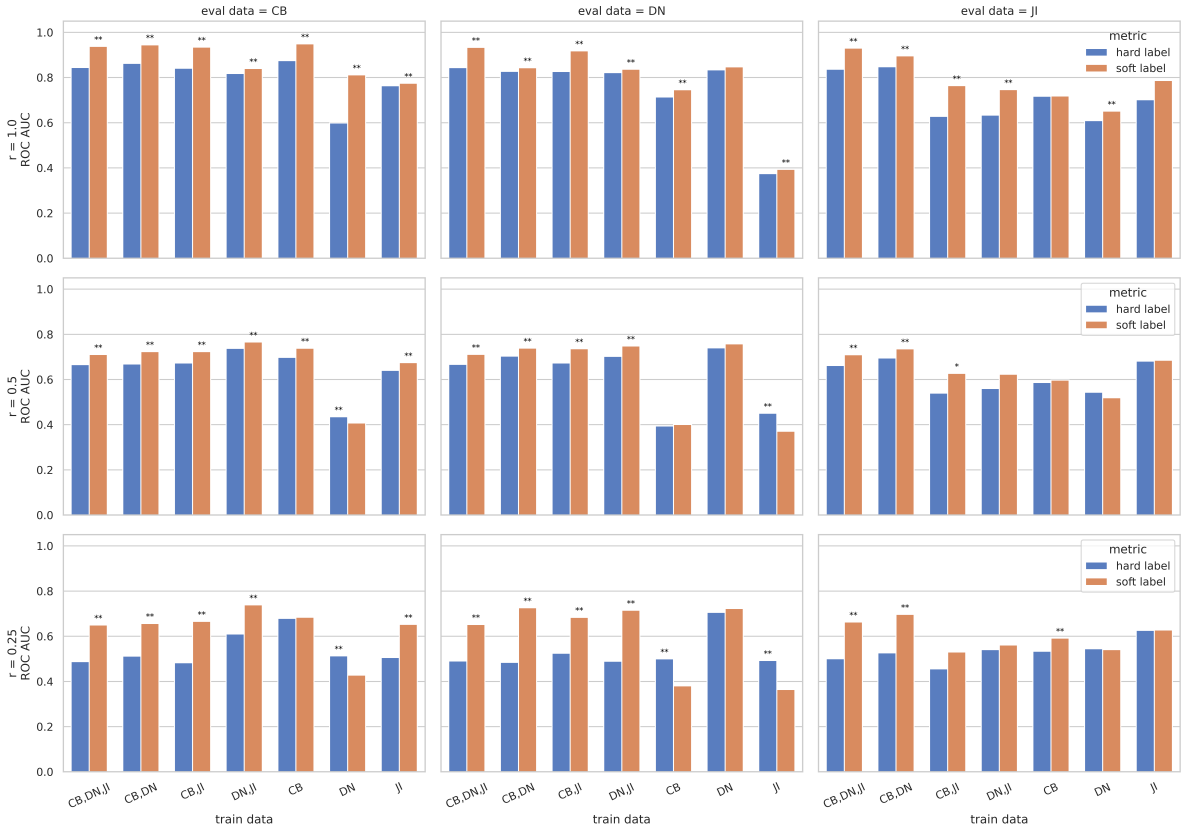

Figure 1: ROC AUC grid for FDANMD. From top to bottom, the prediction results are shown for negotiation progress rates r of 1.0, 0.5, and 0.25, and from left to right, the prediction results are for CB, DN, and JI. Each graph shows the ROC AUC values obtained by performing FDANMD on seven patterns of training data: CB only, DN only, JI only, CB and DN, CB and JI, DN and JI, and CB, DN, and JI. Blue (hard label) shows the ROC AUC values obtained from rule-based annotated hard labels, and orange (soft label) shows the soft labels obtained with ZeSTEM. The * above the bars indicates $p < 0.05$, and ** indicates $p < 0.01$, showing that a statistically significant difference between using annotated hard labels and using soft labels by ZeSTEM was observed.

was particularly noticeable in anticipation from negotiation dialogues when the negotiation progress rate was $r = 0.25$, indicating an early stage of negotiation.

Since soft labels can be obtained zero-shot and general-purpose predictors can be trained inexpensively, we expect this to lead to the construction of LLM-based automated negotiation dialogue agents that can evaluate negotiation situations in various domains and provide appropriate support.

Limitations

The superiority of soft labels over hard labels was confirmed only in predicting whether negotiations would be successful, so we have not confirmed their usefulness in predicting other factors that should be considered during negotiations, such as final profits or satisfaction with the outcome.

Another limitation is the granularity of the dialogue-act inventory. The current labels are useful for controlled comparison with existing rule-based annotations, but they collapse several negotiation-specific moves into broad categories. For example, initial offers and counter-offers are both treated as proposals, soft refusals and hard refusals are not distinguished, and conditional acceptance is not separated from simple agreement. Future work should investigate whether ZeSTEM can be extended to richer negotiation-specific speech-act inventories, including offer types, concession moves, pressure framed as inquiry, relational repair, and implicit rejection.

We also did not conduct human evaluation of the soft labels. Although comparisons with rule-based labels and qualitative examples suggest that ZeSTEM can capture some context-dependent com-

municative functions, rule-based labels are not always a reliable gold standard. Future work should ask human annotators to evaluate whether the top-ranked and lower-ranked soft-label dimensions reflect plausible communicative functions in context.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 23K24897, 24K22333 and JST FOREST (Fusion Oriented REsearch for disruptive Science and Technology) Grant Number JPMJFR216S and JST SPRING (Support for Pioneering Research Initiated by the Next Generation) Grant Number JPMJSP2116.

References

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford [Eng.].

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. 2020. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. 2021. Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 248–257, New York, NY, USA. Association for Computing Machinery.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv e-prints*, pages arXiv–2311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2023. Language of bargaining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13161–13185, Toronto, Canada. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3146–3150, Portorož, Slovenia. European Language Resources Association (ELRA).

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Franz Király, Markus Löning, Tony Bagnall, Matthew Middlehurst, Anirban Ray, Sajaysurya Ganesh, Martin Walter, George Oastler, Jason Lines, ViktorKaz, Benedikt Heidrich, Lukasz Mentel, Jigyasu, Sagar Mishra, chrisholder, Daniel Bartling, Armaghan Shakir, Leonidas Tsaprounis, RNKuhns, and 10 others. 2025. sktime/sktime: v0.40.1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model

serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J. Király. 2019. sktime: A unified interface for machine learning with time series. *Preprint*, arXiv:1909.07872.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Johannes Schneider, Steffi Haag, and Leona Chandra Kruse. 2023. Negotiating with llms: Prompt hacks, skill gaps, and reasoning deficits. *arXiv e-prints*, pages arXiv–2312.

John R. Searle. 1975. Indirect speech acts. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, volume 3, pages 59–82. Academic Press, New York.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Ken Watanabe and Katsuhide Fujita. 2025. Fine-tuning models for final disagreement anticipation in negotiation mid-dialogues. *IEICE Transactions on Information and Systems*, E108.D(3):286–294.

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.

Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. 2021. Dialogue act-based breakdown detection in negotiation dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 745–757, Online. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378, Stockholm, Sweden. Association for Computational Linguistics.

A Detailed Texts input into the Text Embedding Models

A.1 Variations of Converting Label Information into Text

Three methods for converting label information into text for input into the model are shown in Table 5.

A.2 Variations of Instructions

Three patterns of instructions to be input along with the utterance to be soft-labeled are shown in Table 6.

Table 5: Three methods for converting label information into text. In Methods, Word indicates the label word itself, Simple indicates a simple sentence, and Detailed indicates a detailed sentence with an explanation.

Methods	Dialogue Acts	Text
Word	<greet>	Greet
	<inquire>	Inquire
	<inform>	Inform
	<propose>	Propose
	<agree>	Agree
	<disagree>	Disagree
	<unk>	Unknown
Simple	<greet>	"The dialogue act of the statement is greet."
	<inquire>	"The dialogue act of the statement is inquire."
	<inform>	"The dialogue act of the statement is inform."
	<propose>	"The dialogue act of the statement is propose."
	<agree>	"The dialogue act of the statement is agree."
	<disagree>	"The dialogue act of the statement is disagree."
	<unk>	"The dialogue act of the statement is unknown."
Detailed	<greet>	"The dialogue act of the statement is greet, which includes words like hi, hello, yo, hey, hiya, howdy, how are you, good day, good afternoon and good morning."
	<inquire>	"The dialogue act of the statement is inquire, which includes words like what,where, when, which, how's, how about, how does, do you, did you, will you, would you, could you, are you, do we, did we, could we, do i, let me know and ?."
	<inform>	"The dialogue act of the statement is inform, which includes sentences a previous utterance ends with <inquire> and its reply does not contain any other tags."
	<propose>	"The dialogue act of the statement is propose, which includes any digits, words like come down, highest, lowest, go higher/lower and i would like, or a new intermediate offer is proposed."
	<agree>	"The dialogue act of the statement is agree, which includes words like ok, okay, no problem, yes, great, perfect, thanks, gracias, thx, thank you, pleasure, f ine, deal, cool, that works, that will work, that works, it will work, sounds good, very good, looks good and i can do."
	<disagree>	"The dialogue act of the statement is disagree, which includes words like isn't, worse, bad, sorry, no, not, nothing, don't, can't, cannot, afraid, a lot lower/higher and too much/high/low, or an intermediate offer is rejected."
	<unk>	"The dialogue act of the statement is unknown."

Table 6: Instructions input along with the utterances. Genre indicates that instructions include only the genre to be labeled, Word indicates that instructions include the list of label words, and Detailed indicates that instructions include the list of labels with detailed descriptions.

Patterns	Instructions
Genre	"Given a statement in a dialogue, classify the statement into which dialogue act it belongs."
Word	"Given a pair ofstatements in a dialogue, classify the current statement as belonging to one of the following dialogue acts: Greet, Inquire, Inform, Propose, Agree, Disagree, or Unknown."
Detailed	"Given a pair ofstatements in a dialogue, classify the current statement as belonging to one of the following dialogue acts: - Greet; which includes words like hi, hello, yo, hey, hiya, howdy, how are you, good day, good afternoon and good morning. - Inquire; which includes words like what,where, when, which, how's, how about, how does, do you, did you, will you, would you, could you, are you, do we, did we, could we, do i, let me know and ?. - Inform; which includes sentences where the previous utterance ends with Inquire and its reply does not contain any other tags. - Propose; which includes any digits, words like come down, highest, lowest, go higher/lower and i would like, or a new intermediate offer is proposed. - Agree; which includes words like ok, okay, no problem, yes, great, perfect, thanks, gracias, thx, thank you, pleasure, f ine, deal, cool, that works, that will work, that works, it will work, sounds good, very good, looks good and i can do. - Disagree;which includes words like isn't, worse, bad, sorry, no, not, nothing, don't, can't, cannot, afraid, a lot lower/higher and too much/high/low, or an intermediate offer is rejected. - Unknown; none of the above applies."

B Numerical Information of Experimental Results of FDANMD using Soft-labels by ZeSTEM

Table 7,8 and 9 show the ROC AUC values obtained by performing FDANMD on seven patterns of training data in three different negotiation progress rates. Each table describes the evaluation of learning using rule-based annotated hard labels and learning using ZeSTEM-based soft labels, as well as the differences in their results and whether or not those differences are statistically significant.

Table 7: ROC AUC values of FDANMD in the negotiation progress rate $r = 1.0$. The * in Significance indicates $p < 0.05$, and ** indicates $p < 0.01$, showing that a statistically significant difference between using annotated hard labels and using soft labels by ZeSTEM was observed.

Eval Data	Train Data	Hard Label	Soft Label	Difference (Soft - Hard)	Significance
CB	CB,DN,JI	0.844657	0.937893	0.093236	**
CB	CB,DN	0.862574	0.944891	0.082317	**
CB	CB,JI	0.840930	0.934714	0.093784	**
CB	DN,JI	0.817999	0.840081	0.022082	**
CB	CB	0.874799	0.948819	0.074020	**
CB	DN	0.598696	0.811699	0.213003	**
CB	JI	0.763793	0.774664	0.010871	**
DN	CB,DN,JI	0.843804	0.933330	0.089526	**
DN	CB,DN	0.827311	0.843535	0.016224	**
DN	CB,JI	0.827049	0.917967	0.090918	**
DN	DN,JI	0.821603	0.836814	0.015211	**
DN	CB	0.714126	0.746087	0.031961	**
DN	DN	0.833845	0.846998	0.013153	
DN	JI	0.374459	0.393730	0.019271	**
JI	CB,DN,JI	0.836574	0.930285	0.093711	**
JI	CB,DN	0.847656	0.895689	0.048033	**
JI	CB,JI	0.628175	0.764217	0.136042	**
JI	DN,JI	0.633902	0.746572	0.112670	**
JI	CB	0.717344	0.718393	0.001049	
JI	DN	0.609177	0.651267	0.042090	**
JI	JI	0.701756	0.786993	0.085237	

Table 8: ROC AUC values of FDANMD in the negotiation progress rate $r = 0.5$. The * in Significance indicates $p < 0.05$, and ** indicates $p < 0.01$, showing that a statistically significant difference between using annotated hard labels and using soft labels by ZeSTEM was observed.

Eval Data	Train Data	Hard Label	Soft Label	Difference (Soft - Hard)	Significance
CB	CB,DN,JI	0.665833	0.710406	0.044573	**
CB	CB,DN	0.668829	0.723507	0.054678	**
CB	CB,JI	0.673100	0.723574	0.050474	**
CB	DN,JI	0.737733	0.766228	0.028495	**
CB	CB	0.697760	0.738576	0.040816	**
CB	DN	0.435287	0.407240	−0.028047	**
CB	JI	0.640208	0.674484	0.034276	**
DN	CB,DN,JI	0.667038	0.711252	0.044214	**
DN	CB,DN	0.703504	0.738871	0.035367	**
DN	CB,JI	0.672828	0.736251	0.063423	**
DN	DN,JI	0.702299	0.748389	0.046090	**
DN	CB	0.394036	0.400589	0.006553	*
DN	DN	0.739742	0.757445	0.017703	
DN	JI	0.450702	0.371027	−0.079675	**
JI	CB,DN,JI	0.662168	0.709358	0.047190	**
JI	CB,DN	0.695403	0.735453	0.040050	**
JI	CB,JI	0.539708	0.627067	0.087359	*
JI	DN,JI	0.560082	0.622976	0.062894	
JI	CB	0.586884	0.597021	0.010137	
JI	DN	0.543292	0.519216	−0.024076	
JI	JI	0.681099	0.685425	0.004326	

Table 9: ROC AUC values of FDANMD in the negotiation progress rate $r = 0.25$. The * in Significance indicates $p < 0.05$, and ** indicates $p < 0.01$, showing that a statistically significant difference between using annotated hard labels and using soft labels by ZeSTEM was observed.

Eval Data	Train Data	Hard Label	Soft Label	Difference (Soft - Hard)	Significance
CB	CB,DN,JI	0.487794	0.649896	0.162102	**
CB	CB,DN	0.512294	0.656776	0.144482	**
CB	CB,JI	0.482876	0.665936	0.183060	**
CB	DN,JI	0.609979	0.738825	0.128846	**
CB	CB	0.679283	0.684135	0.004852	
CB	DN	0.512935	0.427762	−0.085173	**
CB	JI	0.505921	0.652703	0.146782	**
DN	CB,DN,JI	0.490305	0.652019	0.161714	**
DN	CB,DN	0.484171	0.726112	0.241941	**
DN	CB,JI	0.524656	0.683885	0.159229	**
DN	DN,JI	0.489277	0.715244	0.225967	**
DN	CB	0.499906	0.380497	−0.119409	**
DN	DN	0.705791	0.723085	0.017294	
DN	JI	0.492694	0.364154	−0.128540	**
JI	CB,DN,JI	0.500723	0.663259	0.162536	**
JI	CB,DN	0.526462	0.696797	0.170335	**
JI	CB,JI	0.455730	0.530323	0.074593	
JI	DN,JI	0.540821	0.561564	0.020743	
JI	CB	0.533578	0.591757	0.058179	**
JI	DN	0.544697	0.540914	−0.003783	
JI	JI	0.626339	0.627516	0.001177	

C Comparisons with Prior Study Methods and Cutting-edge Text-based LLMs

Figure 2, 3 and 4 show the performance comparison in ROC AUC scores with prior study methods and cutting-edge text-based LLMs. From prior study methods by Watanabe and Fujita (2025), we choose text-based BERT-based models and dialogue acts-based GRU-based models for comparison, both models trained only on target dataset and models pretrained on CB and then fine-tuned on the target dataset. As cutting-edge LLMs, we choose GPT-5 mini and Gemini 2.5 Flash, closed models that requires access via API. In addition, as open-weight LLMs that can run on one RTX 5090, we choose gpt-oss-20b, Magistral Small 1.2, Qwen3 30B A3B 2507 and Nemotron 3 Nano 30B A3B. For all LLMs, negotiation dialogues for anticipating final disagreement are given in text format and FDANMD was performed in zero-shot.

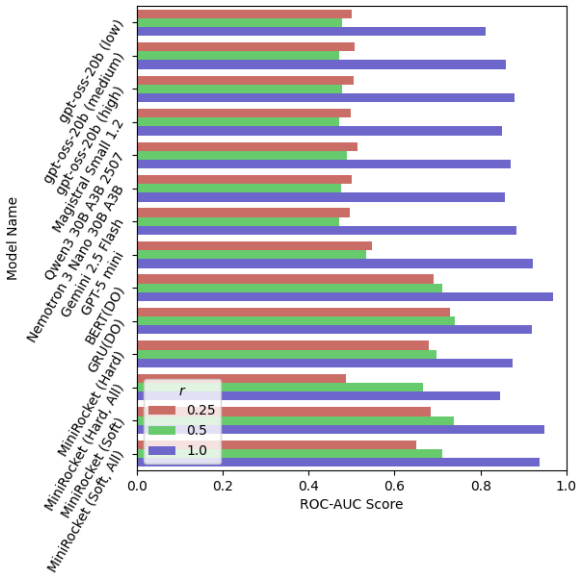

Figure 2: Performance comparison results for FDANMD in CB. r indicates the negotiation progress rate. (DO) refers to a model trained only on the target dataset, (Hard) refers to a model trained on hard labels, (Soft) refers to a model trained on soft labels and (All) refers to a model trained on all of CB, DN and JI.

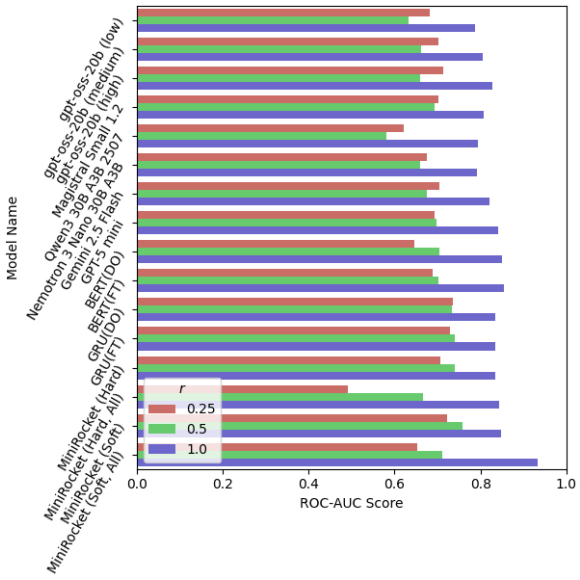

Figure 3: Performance comparison results for FDANMD in DN. r indicates the negotiation progress rate. (DO) refers to a model trained only on the target dataset, (FT) refers to a model pretrained on CB and then fine-tuned on the target dataset, (Hard) refers to a model trained on hard labels, (Soft) refers to a model trained on soft labels and (All) refers to a model trained on all of CB, DN and JI.

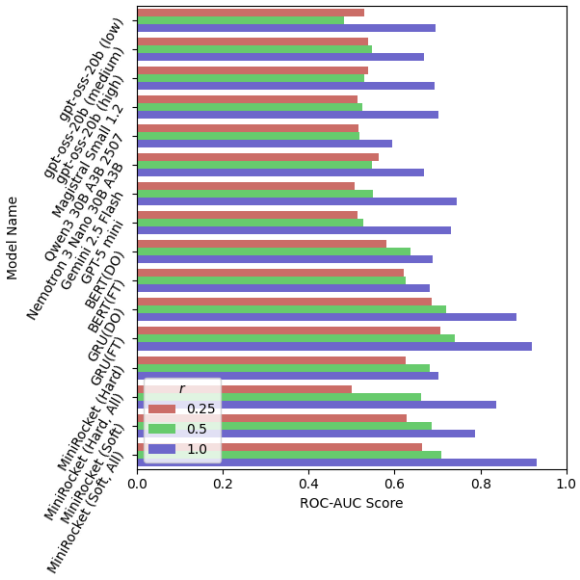

Figure 4: Performance comparison results for FDANMD in JI. r indicates the negotiation progress rate. (DO) refers to a model trained only on the target dataset, (FT) refers to a model pretrained on CB and then fine-tuned on the target dataset, (Hard) refers to a model trained on hard labels, (Soft) refers to a model trained on soft labels and (All) refers to a model trained on all of CB, DN and JI.