# Investigating Dialogue Act Classification through Cross-Corpora Fine-Tuning of Pretrained Language Models

Anonymous ACL submission

## Abstract

Fine-tuning pre-trained language models (PLMs) has achieved significant performance improvements in natural language understanding tasks such as dialogue act classification. However, most of these models are evaluated and benchmarked on standard datasets, and often do not perform well in practical, real-world scenarios such as our scenario of interest: dialogues of collaborative human learning, in which two learners work together to solve a problem in a classroom. To address this challenging scenario, we fine-tuned variants of the RoBERTa and LLaMA-2 models for dialogue act classification within using cross-corpora model fine-tuning approaches on two corpora of collaborative learning dialogues. Our experiments show that fine-tuning PLMs using cross-corpora approaches has the potential to improve classification performance, especially when a corpus has limited representation of certain dialogue acts. This work highlights the potential of using this approach for future domain-specific dialogue act classification tasks.

## 1 Introduction

In recent years, pretrained language models (PLMs) have revolutionized the field of natural language processing, largely driven by their improved performance when fine-tuned for downstream tasks (Devlin et al., 2019; Yang et al., 2019; Radford et al., 2019; Raffel et al., 2020; Zhuang et al., 2021; Touvron et al., 2023b). PLMs are now serving as an effective starting point for many NLP tasks, including dialogue act classification (Saha et al., 2019b; Li et al., 2020). However, most of these models are evaluated and benchmarked on standard dialogue datasets, and they may not perform well in real-world scenarios. Our scenario of interest is dialogue act classification in the context of collaborative learning, in which two learners work together to solve a problem within a classroom.

In the context of collaborative learning, dialogue acts serve important modeling goals: they represent the pragmatics of utterances, offer many cues for assessing the effectiveness of collaboration, and provide insight into the kinds of dialogue behaviors that impact learning, performance, and problem-solving abilities (Chi and Wylie, 2014; Borge et al., 2019; Snyder et al., 2019). There is a great need to leverage the robust capabilities of PLMs for dialogue acts classification on collaborative learning dialogue corpora.

With that in mind, our work investigates the results of fine-tuning RoBERTa (Zhuang et al., 2021), known for state-of-the-art performance in classification tasks, and the recently released open-sourced LLaMA-2 (Touvron et al., 2023b), which to the best of our knowledge has not been used for dialogue act classification. However, fine-tuning these PLMs for dialogue act classification within the context of collaborative learning is challenging for several reasons: the availability of annotated collaborative learning dialogue datasets is highly limited, and there is high domain-specificity of the language present in these corpora.

To address these challenges, we apply a cross-corpora approach, which involves leveraging the properties of one dataset to improve the performance of another related dataset (Webb and Ferguson, 2010). We conducted experiments using the cross-corpora fine-tuning approaches on both RoBERTA and LLaMA-2 models to investigate the following research questions: ***RQ1)*** *How does fine-tuning PLMs using cross-corpora fine-tuning impact the performance of dialogue act classification in the context of collaborative learning?*, and ***RQ2)*** *What dialogue acts can be learned from one domain to another during cross-corpora fine-tuning?*

The novel contributions of this work are as follows: (1) We provide evidence for the application of cross-corpora fine-tuning approaches for dialogue act classification; and (2) We show results of cross-corpora fine-tuned models outperforming baselines in scenarios with limited dialogue act representation.

## 2 Experimental Settings

### 2.1 Annotated Corpora

We evaluate the cross-corpora approach on two collaborative dialogue corpora.

**Corpus I:** A transcribed spoken corpus comprising 6205 utterances from 18 video recordings of 36 paired middle school learners engaged in collaborative coding activities. Its 15 dialogue acts were manually annotated with a Cohen's kappa of 0.83.

**Corpus II:** A textual corpus comprising 3401 utterances from the textual chats of 68 sessions of 136 paired high school learners during collaborative computational music remixing. It features 16 fine-grained dialogue acts manually annotated with a Cohen's kappa 0.76.

The original labeling of each corpus produced sparsity issues for some dialogue acts, so we mapped the original labeling onto six main classes (Table 1).

| DA class | # of samples (*Corpus I*) | # of samples (*Corpus II*) |
|---|---|---|
| QUESTION | 1107 | 83 |
| RESPONSE | 331 | 580 |
| STATEMENT | 2555 | 937 |
| FEEDBACK | 203 | 258 |
| SUGGESTION | 809 | 1023 |
| OTHER | 865 | 520 |

Table 1: Distribution of the DA classes Across the Two Datasets

For cross-corpora training and testing, we split the datasets into approximately 80/20 splits stratifying by pairs and labels to ensure a proportional distribution of the dialogue acts and that no individual paired session's utterances were present in both training and test set.

For our experiments, we fine-tuned two PLMs, RoBERTa$_{base}$ and the LLaMA-2 7B model.

**RoBERTa** (Zhuang et al., 2021), Robustly Optimized BERT approach, builds on the original BERT and modifies the pretraining strategies, such as using byte-pair encoding (Kudo and Richardson, 2018; Wang et al., 2020), modifying BERT's static MLM objective to a dynamic MLP, removing the next-sentence pretraining objectives and modifying key training parameters. Recently, RoBERTa has been found to outperform other traditional deep learning and BERT models for DA classification tasks (Duran et al., 2023)

**LLaMA-2** (Touvron et al., 2023b) is a collection of newly released open-source LLMs based on the LLaMA (Touvron et al., 2023a) by Meta GenAI. The release of these open-source LLaMA-2 models creates opportunities for the research community to fine-tune the actual weights and biases of the models with transparency and visibility to the model architecture and pretraining process. However, like most recent LLMs, LLaMA-2 is a decoder-only transformer model developed mainly for generative tasks. As such, we used a crossentropy loss function (Equation 1) between neural model's output logits and target labels, averaged over the mini-batch size $N$ for the classification task where $log(p_{n,c})$ is the natural logarithm of the predicted probability that observation $n$ belongs to class $c$, $y_{n,c}$ is the binary indicator for the true class label for each sample $n$ and class $c$.

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c}) \quad (1)$$

***4-bit Quantization of LLaMA-2***: Despite the open access to LLaMA-2 models, the high computational demands pose significant challenges. For instance, fine-tuning a LLaMA-2 7B model requires approximately 112GB of GPU memory, exceeding the capacity of consumer GPUs. To mitigate this, there has been a growing interest in parameter efficient fine-tuning (PEFT) (Houlsby et al., 2019) quantization approaches. Recently, 4-bit quantization has shown optimal performance resulting in reduced latency and memory use (Dettmers and Zettlemoyer, 2023). Equation 2 shows the formula for quantizing a 32-bit Floating Point (FP32) tensor into a Int4 tensor with magnitude of [-7,7].

$$X_{Int4} = round\left(\frac{7}{absmax(X_{FP32})} \times X_{FP32}\right) \quad (2)$$

### 2.2 Baseline models

We considered six baselines for our experiments, where each baseline involved fine-tuning a PLM on a specific corpus and evaluating it using the same corpus's test set. We trained and evaluated RoBERTa base and LLaMA-2 7B models on each corpus. Additionally, to evaluate the perfomance associated with data balancing, we trained and evaluated additional RoBERTa models with datasets augumented by *upsampling* using the Random Oversampling (ROS) approach (Mohammed et al., 2020).

### 2.3 Cross-corpora fine-tuned models

For our cross-corpora fine-tuning, we conducted experiments based on three cross-corpora conditions as follows:

- **Roberta$_{base}$/ LLaMA-7B_(C1+C2)** : We fine-tuned using a combination of Corpus I and Corpus II training set.
- **(Roberta$_{base}$/ LLaMA-7B_(C1)→C2** : We fine-tuned using Corpus I training set, then fine-tuned the resulting model on Corpus II training set.
- **Roberta$_{base}$/ LLaMA-7B_(C2)→C1** : We fine-tuned using Corpus II training set, then fine-tuned the resulting model on Corpus I training set.

All implementation is done in PyTorch (Paszke et al., 2019). For each fine-tuning experiment with both variants of RoBERTa, we set the following hyperparameters: we used a batch size of 16 with an AdamW Optimizer with a learning rate of 1 e-5 and weight decay of 0.01. We trained for 20 epochs, with early stopping set at 10. For LLaMA-2 7B model, we use the *bitandbytes* (Dettmers et al., 2022) library for the model quantization configuration. We attempted to use QLoRA (Dettmers et al., 2023) with LoRA (Hu et al., 2021), which enabled us to fine-tune only about 1% of the parameters, but we faced the challenge of fine-tuning the LoRA and the 4bit model with the second corpus for our cross-corpora fine-tuning approaches, so we used the QLoRA configuration without LoRA, fine-tuning about 3.9% of the parameters. We trained the quantized model for 10 epochs using a batch size of 4 with a learning rate of 2 e-4 and a maximum sequence length of 512. To save memory, we use a paged 32-bit AdamW optimizer(Kingma and Ba, 2014) and weight decay of 0.05 and mixed precision (Micikevicius et al., 2017). All training was done using single NVIDIA A100 GPU.

## 3 Results and Discussion

To evaluate the performance of the fine-tuned models, we report the overall Accuracy and F1 (macro) score, which is the arithmetic mean of individual class F1 scores giving equal weight to all classes (Table 2). We also report macro-averaged recall and F1 scores for each class (Table 3).

Table 2: DA classification results (accuracy (Acc) and F1 scores) for fine-tuned models compared to baselines.

| *Datasets* | Corpus I *(C1)* | | Corpus II *(C2)* | |
|---|---|---|---|---|
| **Model** | **Acc** | **F1** | **Acc** | **F1** |
| **Baseline models** | | | | |
| RoBERTa_c1 | 0.750 | 0.678 | - | - |
| RoBERTa_c2 | - | - | 0.684 | 0.587 |
| *ups*_RoBERTa_c1 | 0.686 | 0.622 | - | - |
| *ups*_RoBERTa_c2 | - | - | 0.677 | 0.625 |
| LLaMA-2-7B_c1 | 0.697 | 0.609 | - | - |
| LLaMA-2-7B_c2 | - | - | 0.669 | 0.600 |
| **Cross-corpora models** | | | | |
| RoBERTa_c1+c2 | 0.726 | 0.626 | 0.630 | 0.568 |
| RoBERTa_c1->c2 | - | - | 0.697 | 0.635 |
| RoBERTa_c2->c1 | **0.756** | **0.683** | - | - |
| *ups*_RoBERTa_c1+c2 | 0.719 | 0.631 | 0.643 | 0.586 |
| *ups*_RoBERTa_c1->c2 | - | - | **0.699** | **0.645** |
| *ups*_RoBERTa_c2->c1 | 0.674 | 0.629 | - | - |
| LLaMA-2-7B_c1+c2 | 0.677 | 0.592 | 0.597 | 0.548 |
| LLaMA-2-7B_c1->c2 | - | - | 0.647 | 0.590 |
| LLaMA-2-7B_c2->c1 | 0.716 | 0.628 | - | - |

### 3.1 Comparison to baseline models

To answer RQ1, we compared the results of the cross-corpora fine-tuned models to the baselines, displayed in Table 2. As hypothesized, having more than one domain-related dataset incorporated in the model fine-tuning improves the model performance, especially in the cross-corpora fine-tuning where we learning from a model based on dataset from a similar domain fine-tuned on a PLM. Our cross-corpora approach using RoBERTa had the best performance on the both test sets. Our best performing models achieved an F1 score of 0.683 on the corpus I test set, and an F1 score of 0.645 on the corpus II test set.

### 3.2 Comparison across DA classes

To answer RQ2, we compared the results of the cross-corpora fine-tuned models for each individual DA class to examine which of the DAs were learned across corpora. The main motivation of our approach is that cross-corpora learning helps to improve the recall of dialogue acts that are scarce in a given corpus. Table 3 shows the impact of our cross-corpora approach in improving the recall especially in cases with limited DA available in the corpus. In particular, when we trained the models on a dataset with a larger amount of a give DA, it significantly improves the performance when evaluated on a dataset with a smaller amount of the given DA. Our results show that all models that were fine-tuned with Corpus I, which have a significant amount of *QUE* tags compared to Corpus II, when evaluated on the Corpus 2 data, showed an improvement in detecting the *QUE* tag. Similarly,

| | *Dialogue Acts* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | QUE | | RES | | STMT | | SU | | FDBK | | OTH | |
| **Model** | R | F1 | R | F1 | R | F1 | R | F1 | R | F1 | R | F1 |
| ***Corpus 1*** | | | | | | | | | | | | |
| RoBERTa_c1 | 0.927 | 0.882 | 0.697 | **0.775** | **0.878** | 0.790 | 0.546 | 0.611 | **0.529** | **0.439** | **0.443** | **0.570** |
| LLaMA-2_c1 | 0.873 | 0.859 | 0.708 | 0.663 | 0.794 | 0.747 | 0.596 | 0.598 | 0.353 | 0.316 | 0.371 | 0.469 |
| RoBERTa_c1+c2 | 0.902 | 0.871 | **0.831** | 0.715 | 0.781 | 0.769 | **0.794** | **0.667** | 0.235 | 0.242 | 0.351 | 0.493 |
| LLaMA-2_c1+c2 | 0.824 | 0.837 | 0.742 | 0.695 | 0.785 | 0.736 | 0.546 | 0.535 | 0.294 | 0.312 | 0.356 | 0.438 |
| RoBERTa_c2->c1 | **0.951** | **0.884** | 0.685 | 0.753 | 0.846 | **0.795** | 0.681 | 0.671 | 0.529 | 0.429 | 0.438 | 0.567 |
| LLaMA_c2->c1 | 0.883 | 0.860 | 0.719 | 0.656 | 0.819 | 0.767 | 0.610 | 0.637 | 0.353 | 0.343 | 0.397 | 0.503 |
| ***Corpus 2*** | | | | | | | | | | | | |
| RoBERTa_c2 | 0.167 | 0.242 | **0.767** | 0.671 | **0.697** | **0.670** | 0.824 | 0.784 | 0.561 | 0.547 | 0.468 | 0.608 |
| LLaMA-2_c2 | 0.333 | 0.421 | 0.677 | 0.684 | 0.693 | 0.643 | 0.709 | 0.722 | 0.421 | 0.440 | **0.654** | **0.667** |
| RoBERTa_c1+c2 | 0.500 | 0.245 | 0.744 | 0.669 | 0.550 | 0.584 | 0.745 | 0.746 | **0.561** | **0.587** | 0.487 | 0.578 |
| LLaMA-2_c1+c2 | **0.500** | 0.250 | 0.677 | 0.682 | 0.573 | 0.547 | 0.626 | 0.686 | 0.544 | 0.544 | 0.545 | 0.578 |
| RoBERTa_c1->c2 | 0.417 | **0.476** | 0.759 | 0.692 | 0.638 | 0.653 | **0.856** | **0.795** | 0.544 | 0.530 | 0.545 | 0.664 |
| LLaMA-2_c1->c2 | 0.333 | 0.432 | 0.692 | **0.702** | 0.693 | 0.629 | 0.687 | 0.692 | 0.439 | 0.439 | 0.596 | 0.648 |

Table 3: Evaluation results for individual DA classes on the two collaborative learning corpora for the RoBERTA base and 4-bit quantized LLaMA-2 7B model using cross corpora fine-tuning approach. Recall scores are reported to show the model's ability to correctly identify the actual DAs.

since Corpus II has more *SU* tag, when it is used in cross-corpora fine-tuning and evaluated on Corpus I, the significantly improves the correct detection of the *SU* tag.

## 4 Related Work

In recent years, the introduction of the Transformers architecture by Vaswani et al. (Vaswani et al., 2017) has paved the way for high-performing transformer-based language models, such as BERT (Devlin et al., 2018) and GPT (Floridi and Chiriatti, 2020), which have demonstrated remarkable performance in various NLP tasks. These models have been used on dialogue datasets such as SWBD (Jurafsky, 1997) and MRDA (Shriberg et al., 2004) to establish benchmarks for DA classification models. Due to the improved performance by BERT on DA classification, researchers have also experimented with BERT-based models and compared their performance to the original BERT (Saha et al., 2020a; Qin et al., 2021; Li et al., 2022). As a result of the improved performance of BERT-based models compared to earlier deep learning models like RNN and LSTM, more researchers have applied these models to several real-world datasets such as, speech acts classification with the Twitter corpus (Saha et al., 2020b), achieving SOTA performance of 75.97% on the Twitter dataset collected by (Saha et al., 2019a). More recently, researchers have also explored the robustness of BERT-based models on social media data (Vielsted et al., 2022).

In contrast to encoder-based models like BERT, researchers have also explored decoder-only, like GPT-2, and their potential to perform DA classification, showcasing their extended capabilities (Weng et al., 2020). Recently, researchers have used DialoGPT (Zhang et al., 2019), a dialogue PLM built upon GPT-2, for classifying dialogue acts in K-12 classroom data (Kumaran et al., 2023). With more powerful variants of GPT, such as GPT-3.5 (Floridi and Chiriatti, 2020), GPT-4 (OpenAI, 2023) and LLaMA-2 (Touvron et al., 2023b), there is an increasing opportunity to further the work on DA classification and train better models.

## 5 Conclusion and Future Work

DA classification is an important task in dialogue analysis especially in the context of collaborative learning. However due to insufficient available labeled datasets, it is challenging to train high performing DA classification models. Our work aims to address these challenges by investigating cross-corpora fine-tuning to improve the performance of the models and and to evaluate the ability to better detect DAs in cases where a domain-related corpus has less of a give tag. Further, we explored the newly released LLaMA-2 model for DA classification tasks. We applied quantization to reduce the size of model and fine-tuned a subset of the model parameters, and still achieved on-par results.

In the future, we would like to do perform additional experiments with a third dataset within the same domain, that is not included in the training. We would also like to explore more fine-tuning techniques for DA classification using PLMs. Also, we would consider multi-modal inputs for predicting DAs such as audio and video with the textual inputs.

## 6 Limitations

**Data variability and imbalance:** Our experiments used very similar datasets, however one was speech recordings transcribed to text, making it more robust than the textual interaction data. Students tends to talk more than type. Furthermore, there are some slight difference in the type of dialogue interactions between middle school and high school learners, which can be also reflected in the dialogue. Although these resulted in significant data imbalance, our metaclasses groupings and stratified splitting of the train/test data helped reduce the data imbalance.

**Closed-source data:** Our data is primarily collected from K-12 participants, some of whom are minors, resulting in challenges to sharing our data due to data restrictions and privacy concerns.

**Computational resource limitations:** due to computing limitations, we were unable to investigate the scaling behavior of the LLMs, such investigating with different precision and with larger models like LLaMA-13B. Further experiments and studies are need in the future to investigate the impact of fine-tuning significantly larger PLMs.

## 7 Ethics Statement

Our research work focuses on analyzing dialogue data collected during collaborative learning activities in K-12 settings. For these reasons, the ethical implications of our work include ensuring the privacy of our participants and protection of data collected. Our research studies were conducted with relevant Institutional Review Board (IRB) approval that included written parental consent and student assent. All the researchers involved in the study are trained and certifies on human subject data research, and all the data are stored in dedicated secure machines with restricted access. Our data analysis included the development of DA taxonomy and the annotation of corpora. All annotators were Ph.D. students trained in dialogue act annotation following well-known steps for dialogue act annotation, including the iterative refinement of DA labels and establishing inter-rate agreement (Landis and Koch, 1977). In addition, we recognize that pretrained language models can perpetuate and amplify biases present in training data, and we are cautious of some potential biases during the fine-tuning. We are aware of the environmental impact associated with training large language models. We minimized this impact by efficiently using computational resources and by choosing to fine-tune the larger models using PEFT approaches.

## References


Marcela Borge, Tugce Aldemir, and Yu Xia. 2019. Unpacking socio-metacognitive sense-making patterns to support collaborative discourse.

Michelene TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Nathan Duran, Steve Battle, and Jim Smith. 2023. Sentence encoding for dialogue act classification. *Natural Language Engineering*, 29(3):794–823.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, Snigdha Chaturvedi, and James Lester. 2023. Improving classroom dialogue act recognition from limited labeled data with self-supervised contrastive learning classifiers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10978–10992.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. 2020. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

OpenAI. 2023. Gpt-4 technical report.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13709–13717.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tulika Saha, Srivatsa Ramesh Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. 2020a. Bert-caps: A transformer-based capsule network for tweet act classification. *IEEE Transactions on Computational Social Systems*, 7(5):1168–1179.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020b. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2019a. Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Tulika Saha, Saurabh Srivastava, Mauajama Firdaus, Sriparna Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2019b. Exploring machine learning and deep learning frameworks for task-oriented dialogue act classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Caitlin Snyder, G Biswas, M Emara, S Grover, and L Conlin. 2019. Analyzing students' synergistic learning processes in physics and ct by collaborative discourse analysis. In *Computer-supported collaborative learning*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Marcus Vielsted, Nikolaj Wallenius, and Rob van der Goot. 2022. Increasing robustness for cross-domain dialogue act classification on social media data. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 180–193.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Nick Webb and Michael Ferguson. 2010. Automatic extraction of cue phrases for cross-corpus dialogue act classification. In *Coling 2010: Posters*, pages 1310–1317.

Yue Weng, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, et al. 2020. Joint contextual modeling for asr correction and language understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6349–6353. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.