Dual Process Masking for Dialogue Act Recognition

Yeo Jin Kim[1], **Halim Acosta**[1], **Wookhee Min**[1], **Jonathan Rowe**[1],
Bradford Mott[1], **Snigdha Chaturvedi**[2], **James Lester**[1],

[1]North Carolina State University, [2]UNC Chapel Hill
{ykim32, hacosta, wmin, jprowe, bwmott, lester}@ncsu.edu[1], snigdha@cs.unc.edu[2]

Abstract

Dialogue act recognition is the task of classifying conversational utterances based on their communicative intent or function. To address this problem, we propose a novel two-phase processing approach called Dual-Process Masking. This approach streamlines the task by masking less important tokens in the input, identified through retrospective analysis of their estimated contribution during training. It enhances interpretability by using the masks applied during classification learning. Dual-Process Masking significantly improves performance over strong baselines for dialogue act recognition on a collaborative problem-solving dataset and three public dialogue benchmarks.

1 Introduction

Dialogue act recognition is the task of classifying utterances in a conversation based on their communicative intent (Stolcke et al., 2000). Accurately discerning the intent of each utterance in human-to-human or human-to-machine interactions plays a central role in a broad range of applications such as conversational agents (Kim et al., 2010; Ahmadvand et al., 2019), meeting analysis (Ang et al., 2005), and emotion analysis (Xu et al., 2023). However, defining a universal dialogue act taxonomy is challenging because conversational settings vary widely. Different domains use distinct taxonomies that reflect their specific purposes and characteristics. Hence, automatically capturing the varied linguistic structures associated with dialogue acts requires a large volume of labeled training data that encompasses a wide range of permutations. However, the availability of labeled data in many domains is limited because the labeling process is labor-intensive. There is growing demand for focused analysis of small-scale conversational data across specific domains (e.g., education, science, healthcare, transportation), where specialized taxonomies are often employed (Hu et al., 2022).

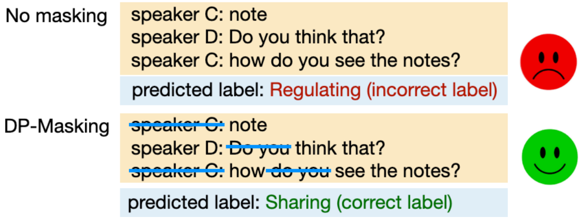

No masking
speaker C: note
speaker D: Do you think that?
speaker C: how do you see the notes?
predicted label: Regulating (incorrect label)
DP-Masking
~~speaker C:~~ note
speaker D: ~~Do you~~ think that?
~~speaker C:~~ how ~~do you~~ see the notes?
predicted label: Sharing (correct label)

Figure 1: No masking vs. Dual-Process Masking (DP-Masking). DP-Masking improves dialogue act recognition by masking less relevant or misleading tokens.

Despite advances in dialogue act recognition via deep learning (Raheja and Tetreault, 2019; Li et al., 2019; Colombo et al., 2020) and large language models (LLMs) (Liu et al., 2019; Noble and Maraev, 2021; Suresh et al., 2021), the effects of limited data and specific tokens on performance remain understudied. In particular, existing approaches based on deep learning, including LLMs, achieve high recognition rates on large-scale data, but their low interpretability limits their applicability to other domains and different dialogue act taxonomies. To bridge this gap, we propose *Dual-Process Masking (DP-Masking)*, which uses masking to de-emphasize tokens that are less relevant to dialogue act recognition and uncover the functional language structure indicative of dialogue acts. As shown in Figure 1, DP-Masking reduces noise by identifying essential structures, thereby enhancing dialogue act recognition accuracy in datasets with limited data. It also improves interpretability by elucidating key tokens crucial for different dialogue acts.

DP-Masking draws from dual-process theories in cognitive psychology (Evans, 2003; Kaufman, 2011; Gronchi and Giovannelli, 2018), which describe two systems of human cognitive processing. The intuitive system is fast, automatic, and instinctive, relying on heuristics and past experiences, while the reflective system is slow, controlled, and rational, requiring deliberate effort for complex

Findings of the Association for Computational Linguistics: EACL 2024, pages 15270–15283
November 12-16, 2024 ©2024 Association for Computational Linguistics

problem-solving (Sun, 2015). These two systems operate in parallel and influence human intelligence and decision-making. In DP-Masking, the reflective system iteratively masks less relevant tokens by learning the tokens' contributions to dialogue acts through controlled effort, while the intuitive system internalizes this learning for rapid, automatic classification.

We validate DP-Masking on a collaborative problem-solving (CPS) dataset collected from student interactions with a collaborative game-based learning environment, as well as with three large-scale public dialogue corpora. This paper primarily focuses on the CPS data because of its limited training set size (∼2K labeled instances). The other datasets are employed to examine DP-Masking performance across varying dataset sizes. For the CPS dataset, DP-Masking achieves state-of-the-art (SOTA) results. We also experiment with different implementations of DP-masking using Llama-2-7B, Llama-3-8B (Touvron et al., 2023), and Flan-T5-small (Chung et al., 2022) as the underlying model and show how DP-Masking outperforms random masking and a no-masking approach, with an average normalized improvement of 13.4% and 17.6% in macro F1 scores, respectively. Our analysis shows how certain token characteristics are associated with masking performance as an interpretation of the learned model and how DP-Masking could lead to performance improvements for dialogue act recognition, highlighting the implications of this research.

For the three public benchmarks, DP-Masking achieves significant improvements with limited data, but its benefit decreases as the data size increases. With the full dataset size, DP-Masking shows performance comparable to the underlying models, while achieving SOTA performance on one of the three benchmarks. Overall, DP-Masking outperforms random masking with an average normalized improvement of 3.2% and a no-masking baseline strategy with an 8.3% improvement in Matthews correlation coefficient (MCC). The contributions of this work include: (1) A novel dialogue act recognition framework, DP-Masking, inspired by dual-process theories of cognition; (2) An evaluation of DP-Masking on a limited-sized educational dataset and three large-scale public benchmarks; (3) An analysis of recognition results to assess their masking strategies.

2 Related Work

2.1 Dialogue Act Recognition

Traditional methods for dialogue act recognition have relied on statistical models such as Hidden Markov models (Stolcke et al., 2000), Bayesian networks (Keizer et al., 2002), and Support Vector Machines (Surendran and Levow, 2006). Subseqeuntly, researchers began to adopt sequential deep learning techniques to better capture temporal dependencies. Recurrent Neural Networks (RNNs) were introduced for this purpose (Bothe et al., 2018), alongside methods that modeled inter-tag dependencies (Kumar et al., 2018; Raheja and Tetreault, 2019; Li et al., 2019), and implemented guided attention mechanisms in Seq2Seq models (Colombo et al., 2020). Some studies focused on incorporating the emotional states of speakers to enhance recognition performance (Saha et al., 2020).

Recent research has utilized large language models (Liu et al., 2019; Noble and Maraev, 2021; Suresh et al., 2021). Most recent work on dialogue act recognition using limited data has explored contrastive learning-based self-supervised approaches using few-shot (Kumaran et al., 2023) and active learning methods based on sample informativeness (Lin et al., 2024). However, the first method requires large volumes of unlabeled data for semi-supervised learning, which is often not readily available, and the second method relies on human annotators for active learning, which contrasts with our goal of automated labeling.

2.2 Dual-Process Theory

Recently, there has been increasing interest in investigating logical analysis, like that in dual-process theories, for language modeling (Goyal and Bengio, 2022). Nye et al. (2021) explored a dual-process-based, no-training approach to address the issue of neural sequence models being fast yet inconsistent. Their findings indicated that infusing logical reasoning inspired by the reflective system bolstered coherence and precision in story creation and following instructions. In a similar vein, Liu et al. (2022) introduced a dual-process theory framework with a neural symbolic processor designed for natural language comprehension. This system excelled in analogical and logical reasoning, outperforming existing competitive techniques in evaluation involving question answering and natural language inference. Our research aims to advance dialogue act recognition by leveraging the synergy between neural networks and logical analysis.

2.3 Language Models and Masking

Masking in pre-training language models involves predicting masked tokens to improve language understanding (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020). Masking techniques are categorized as follows: (a) *causal masking* uses a triangular matrix to prevent the decoder from observing future tokens (Devlin et al., 2019; Raffel et al., 2020); (b) *dynamic masking* learns selective binary masks for pretrained weights in network layers (Zhao et al., 2020), learns mask weights using a soft gate with the sigmoid function (Fan et al., 2021), or jointly masks token n-grams with high collocation (Levine et al., 2021); (c) *random masking* involves randomly replacing a certain percentage of tokens in the input text with a special “mask" token, as used in BERT (Devlin et al., 2019).

DP-Masking differs from prior work that used masking techniques for estimating masked tokens during pre-training. Instead, we focus on selectively masking tokens to capture the functional essence of utterances for specific tasks. To the best of our knowledge, DP-Masking is the first to utilize masking for this purpose. Given our focus on semantic token elimination, positional masking methods such as causal masking (Devlin et al., 2019; Raffel et al., 2020) and position-level dynamic masking (Fan et al., 2021; Zhao et al., 2020) are deemed unsuitable as they focus on word position rather than semantics. In contrast, random masking serves as a viable competitive baseline in this work.

3 Problem Definition

Given an input dialogue, U, consisting of a sequence of utterances, $U = (u_1, u_2, ..., u_j)$, our goal is to output a dialogue act label, $l_i \in \mathcal{L}$, for every utterance u_i. Each utterance is represented as a sequence of tokens: $u_i = (w_1, w_2, ..., w_k)$ where $w \in R^d$ and d is a dimensional token space. The input in our experiments consists of the target utterance, u_i, and k preceding utterances $(u_{i-k}, ..., u_{i-1})$ as context. However, for simplicity, we refer to the concatenation of context and target utterance as an utterance.

4 Dual-Process Masking

Inspired by dual-process theories of cognition, we implement a dual-process masking learning algorithm where the intuitive learning stage and the reflective learning stage iteratively interact and influence dialogue act recognition. Figure 2 shows the overall workflow of DP-Masking. The reflective learning stage actively identifies mask tokens that lead to correct dialogue act labels, and the intuitive learning stage consolidates this knowledge into long-term memory for classification. The core components are a masker M and classifier C pair for each of the reflective learning R and intuitive learning I stages, and a working memory, W, connecting these two stages.

In the reflective learning stage, target masks, which a masker aims to output, are initialized with random masks in the utterances, and this stage's masker M_R learns these target masks using the training data of unmasked utterances. Then, the reflective classifier C_R learns to classify utterances masked by M_R into dialogue act labels. The classification results are analyzed, and the successful masks (i.e., masks from the correctly classified utterances) are updated in the Working Memory W. In the next iteration, the target masks are replaced with successful masks and supplemented with random masks for exploration. In the intuitive learning stage, the intuitive masker M_I learns only from successful masks stored in W, and the intuitive classifier C_I learns classification using utterances masked by M_I.

Algorithm 1 outlines the learning algorithm for DP-Masking. Here, we employ a pretrained language model as the underlying model for each masker and classifier. First, W is initialized by copying training data, U^{tr} (line 1). The algorithm goes through multiple epochs of reflective learning and intuitive learning stages (line 2). In the reflective learning stage, we define three key steps: *exploration*, *analysis*, and *information acquisition*. For *exploration*, $p\%$ of the tokens in U^{tr} (stored in W) are masked randomly and then U^{tr} is randomly split into reflective sub-training (U^{r_tr}) and

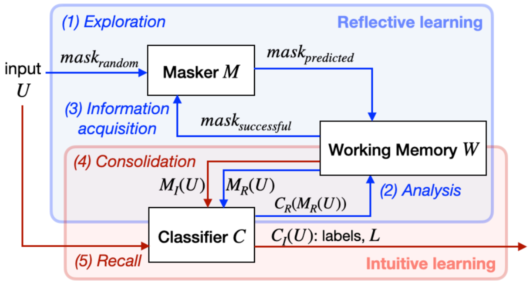


Figure 2: Overall workflow of dual-process masking.

Algorithm 1 Dual-Process Masking learning

```
Input: input $\mathcal{U}$ : $[U^{tr}, U^{v}, U^{te}]$, label $\mathcal{L}$ : $[L^{tr}, L^{v}, L^{te}]$
Parameter: masking rate $p$, max training iteration $e_{max}$
Output: class $c \in \mathcal{C}$
 1: $W \leftarrow$ initializeWorkingMemory($U^{tr}$)
 2: for $e \in [0, e_{max}]$ do
 3:     # Reflective learning stage
 4:     $W \leftarrow$ addRandomMask($W, p$)
 5:     $U^{r\_tr}, L^{r\_tr}, U^{r\_v}, L^{r\_v} \leftarrow$ randomSplit($U^{tr}, L^{tr}$)
 6:     $M_R \leftarrow$ trainMasker($U^{r\_tr}, L^{r\_tr}$)
 7:     $C_R \leftarrow$ trainClassifier($M_R(U^{r\_tr}), L^{r\_tr}$)
 8:     $eval \leftarrow$ evaluate($C_R(M_R(U^{r\_v})), L^{r\_v}$))
 9:     $W \leftarrow$ updateSuccessMasks($eval$)
10:     # Intuitive learning stage
11:     $M_I \leftarrow$ trainMasker($W, L^{tr}$)
12:     $C_I \leftarrow$ trainClassifier($M_I(U^{tr}), L^{tr}$)
13:     $eval \leftarrow$ evaluate($C_I(M_I(U^{v}), L^{v}$))
14:     if $e = e_{max}$ or early stop with $eval$ then
15:         $eval \leftarrow$ evaluate($C_I(U^{te}), L^{te}$)
16:         break
17:     end if
18: end for
```

sub-validation (U^{r_v}) data (line 4-5). The reflective masker, M_R, is trained using U^{r_tr} and the respective labels L^{r_tr} (line 6). The loss function for masker is a cross-entropy cost function that compares predicted masks with target masks. Then, a reflective classifier C_R is trained using masked data $[M_R(U^{r\_tr}), M_R(U^{r\_v})]$ and their respective labels $[L^{r\_tr}, L^{r\_v}]$ (line 7). For *analysis*, the classification results are evaluated from U^{r_v} (line 8), and W is updated with the successful masks for the sub-validation set for *information acquisition* in the next epoch (line 9). Here, target masks are updated with the mask tokens from instances that have been successfully classified as well as new random masks. See Appendix A for an example of updating successful masks using W.

Next, the intuitive learning stage consists of two key steps: *consolidation* and *recall*. In the *consolidation* step, the intuitive masker, M_I, trained with the updated W with successful masks, applies masks to a given text, U_i, to produce the masked input $M_I(U_i)$ (line 11). The intuitive classifier C_I is trained on $M_I(U_i)$ alongside its corresponding target label l_i (line 12). This iterative process reduces the influence of masked tokens in C_I's network, effectively consolidating relevant information into long-term memory. Finally, we consider the training process to be complete when the training epoch caps or meets the early stopping criterion. For inference, the final C_I classifies the unmasked test data U^{te} by *recalling* the learned knowledge to generate evaluation results (line 13-17). It should be noted that during inference, C_I processes unmasked utterances, intuitively assessing token relevance and importance, bypassing M_I. This approach deviates from typical masked language models that predict masked tokens to enhance overall language understanding. Instead, it selectively masks tokens based on their relevance to specific downstream NLP tasks, facilitating more contextual analyses[1].

The rationale for employing distinct masker and classifier pairs for reflective learning and intuitive learning is that the reflective learning models are inherently more susceptible to overfitting, which tends to occur more rapidly during the process of exploration, analysis, and information acquisition on the training data. In contrast, the intuitive learning models eliminate the need for this cycle by using successful masks from W, thereby effectively mitigating overfitting. Since DP-Masking involves training four components (M_R, C_R, M_I, C_I), its time complexity is a fixed multiple of the time complexity of a pretrained language model. For the computational costs of the algorithm and hardware system specifications, see Appendix B.

5 Collaborative Problem-Solving Data

In this section, we conduct an experiment using a collaborative problem-solving (CPS) dataset, which exemplifies our target limited data setting.

Data. The CPS dataset consists of dialogues between groups of three to four middle school students collaboratively working in an educational game, ECOJOURNEYS, as part of their science class on ecosystems to solve a mystery centered on an illness spreading among fish in local farms on a fictional island (Mott et al., 2019). Students communicated via an in-game chat interface throughout the game sessions. To annotate this data, educational researchers designed the dialogue act taxonomy based on four key components of CPS behavior: sharing ideas, negotiating ideas, regulating problem-solving activities, and maintaining communication (Von Davier et al., 2017). We expanded this taxonomy by adding 'other' and 'off-task' categories to more comprehensively represent the dialogue acts observed in the chat during middle-grade collaborative problem solving.

Two senior graduate students with expertise in CPS performed the annotation under the supervision of a senior professor in educational research. Initial annotation of 20% of the data demonstrated

[1] We experimented with a classifier using explicitly masked tokens, but the proposed method yielded better results.

a high inter-rater reliability (IRR) with a Cohen's Kappa of 0.81 (Hong et al., 2024), indicating strong agreement (McHugh, 2012). The two annotators then discussed and resolved their disagreements. Afterward, one of the annotators completed annotating the rest of the dataset, following the guideline for coding qualitative data (Syed and Nelson, 2015). After removing non-dialogue messages, the dataset consists of 1,990 utterances. The data collection was conducted with Institutional Review Board (IRB) approval, and the anonymized CPS data is available for research purposes upon request.

Compared methods. The baselines include BERT-base models (110M parameters), including BERT, BERT-LSTM, and BERT-BiLSTM, along with three dialogue act recognition approaches, SGNN (Ravi and Kozareva, 2018), CASA (Raheja and Tetreault, 2019), RoBERTa-base (125M) and RoBERTa-large (355M) (Liu et al., 2019; Suresh et al., 2021), which are publicly available and are state-of-the-art models.[2] We also compare DP-Masking (DP-*) with random masking (R-*) and no-masking. We apply these three masking strategies to three underlying models: Flan-T5-small (80M), Llama-2-7B, and Llama-3-8B. To improve training efficiency given the high fine-tuning costs of Llama models, only the classifier was trained on Llama, using masked inputs from the FT5 masker. All Llama models were fine-tuned using QLoRA (8-bit quantization) (Dettmers et al., 2023). Hyperparameters are described in Appendix C.

Evaluation. We conducted a 4-fold cross-validation with dialogue data from four student groups, using each group as a test set while splitting the remaining groups into 90% for training and 10% for validation. The evaluations were repeated over three random seeds, yielding a total of 12 results per method. We employed accuracy and macro F1 scores for evaluation metrics due to the imbalanced data and also used normalized improvement (the higher, the better) (Marx and Cummings, 2007), defined in Appendix D.

Results. Table 1 shows the classification results on the CPS data. The majority class is "regulating" at 29.4%. The BERT-based models, including RoBERTa, achieved similar performances with about 41% in both macro F1 score and accuracy. SGNN exhibited a 6% higher accuracy than the BERT-based models but had a 13% lower macro F1 score, whereas CASA underperformed compared to BERT. The FT5 and Llama models generally surpassed the others in terms of accuracy but showed a significant gap between their accuracy and macro F1 scores, signaling their vulnerability to imbalanced data.

Model	macro F1	Accuracy	N.Imp.(F1)%
BERT	**41.4 (1.1)**	**42.0 (1.3)**	0
BERT-LSTM	40.0 (1.3)	40.9 (1.4)	-3.4
BERT-BiLSTM	40.9 (1.2)	40.9 (1.2)	-1.2
RoBERTa-base	**40.7 (1.2)**	41.0 (1.3)	-1.7
RoBERTa-large	**40.5 (1.1)**	41.1 (1.1)	-2.2
CASA-RoBERTa	23.9 (2.6)	39.0 (2.6)	-42.3
SGNN	27.1 (1.0)	**47.0 (3.0)**	-34.5
FT5	30.5 (2.2)	50.6 (3.7)	-26.3
R-FT5	∗38.7 (2.1)	∗58.6 (3.3)	-6.5
DP-FT5	§∗**43.6 (2.0)**	§∗**63.3 (2.7)**	3.8
Llama-2-7B	31.1 (1.5)	53.8 (2.1)	-24.9
R-Llama-2-7B	29.5 (1.7)	48.9 (2.6)	-8.7
DP-Llama-2-7B	§∗**44.7 (3.4)**	§∗**65.8 (2.9)**	5.6
Llama-3-8B	37.7 (2.1)	57.6 (3.0)	-8.9
R-Llama-3-8B	∗40.3 (1.4)	∗60.8 (2.5)	-2.7
DP-Llama-3-8B	§∗**46.6 (2.4)**	§∗**67.2 (2.7)**	8.9

Table 1: Results on the CPS test data. The values in parentheses are the standard errors. ∗ and § indicate the statistical significance of the corresponding model compared to the underlying model and the random masking approach, respectively, within each family of LLMs, using a Wilcoxon rank sum test ($p < 0.05$). Normalized improvement (N.Imp.)(↑) is calculated by comparing each approach with BERT in terms of F1 score. For visual comparisons, refer to Figure 7 in Appendix E.

In comparisons of masking approaches on the Llama and FT5 models, random masking led to decreased performance for R-Llama-2-7B compared to Llama-2-7B without masking. However, random masking significantly improved performance for both R-Llama-3-8B and R-FT5. Notably, DP-Masking consistently resulted in balanced improvements in both macro F1 score and accuracy, with DP-Llama-3-8B achieving the highest scores. Despite its smaller size relative to Llama, R-FT5 and DP-FT5 outperform the base Llama models (Llama-2-7B/-3-8B), demonstrating the benefits of both masking approaches. The base Llama models' lower scores may stem from their fine-tuning process, which updates only a tiny fraction of their parameters (0.04-0.06%, 3-4M) with QLoRA, potentially compromising nuanced classification. In

[2] The code for CASA (github.com/macabdul9/CASA-Dialogue -Act-Classifier) and SGNN (github.com/glicerico/SGNN) is provided by a third party. While the code implements their main concepts, the performance may not fully match the original methods.

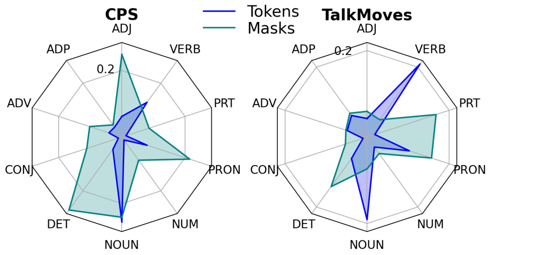


Figure 3: Token and mask distribution by part of speech in the CPS and TalkMoves datasets.

	Masked words
Highest mask frequency rate ($\geq 10\%$)	I'm (43%), I (31%), **submit** (28%), inaudible (27%), **agree** (17%), **screen** (12%), **water** (11.1%)
Lowest mask frequency rate ($< 1\%$)	at (0.04%), in (0.1%), is (0.2%), on (0.3%), ah (0.5%), ey (0.5%), can (0.7%), up (0.7%), **bacteria** (0.8%)

Table 2: Examples of masked words by DP-Masking. The words highlighted in bold are task-related words.

contrast, FT5, a smaller model, can undergo comprehensive fine-tuning for the classification task. This suggests that smaller language models may be more suitable for fine-tuning-based classification tasks with limited resources than their billion-parameter counterparts.

Overall, DP-Masking achieves SOTA results with DP-Llama-3-8B in both macro F1 score and accuracy. Across three strong underlying models (Llama-2-7B, Llama-3-8B, FT5), DP-Masking outperforms random masking with an average normalized improvement of 13.4% and a no-masking strategy with 17.6% in macro F1 score in the CPS dialogue act recognition task.

6 Analysis

We analyze learned masking strategies focusing on 1) masking rate, 2) mask distribution by part of speech, 3) mask types, 4) effect of input length, 5) accuracy by class, and 6) qualitative analysis with successful and failed masking examples.

Masking rate. The masking rate is defined as the fraction of tokens masked in each dataset. For DP-FT5, the masking rates were 13.5% on the CPS dataset (2K) and 6.8% on the larger TalkMoves datasets (170K), described in Section 7. The lower masking rate in TalkMoves suggests that the classifier, which was trained on extensive data, performs well during the iterative training process before the masker has fully learned masking. This variability in masking rates, tailored to the characteristics of the data, contrasts with the fixed 15% masking rate of BERT style models.

Mask distribution. In this experiment, we analyzed part of speech (POS) for token and mask distribution. Figure 3 shows the distribution of all tokens (in blue) and masked tokens (in green). We can see that both CPS and TalkMoves datasets have similar token distributions, with nouns, verbs, and pronouns being predominant. However, their mask distributions differ: CPS has higher masking rates for determiners and articles (DET), nouns (NOUN), and adjectives (ADJ), while TalkMoves exhibits higher masking rates for particles (PRT), pronouns (PRON), and DET. This suggests that although these dialogue datasets share formal linguistic similarities, the importance of tokens varies across different dialogue act taxonomies. Therefore, it is necessary to learn the masking strategy in a data-driven manner instead of random masking. It also explains why DP-Masking outperformed the random baselines in Section 5. We also analyzed the POS distributions of masks by class label (Figure 8 in Appendix E). We made similar observations where each class's POS distributions are different from the CPS and TalkMoves datasets.

Mask types. To examine masked word characteristics, we analyzed the mask frequency rate by word, defined as the frequency of a masked word relative to its total word frequency. Table 2 displays the highest and lowest mask frequency rates of words masked by DP-Masking on the CPS data, with the numbers in parentheses representing the mask frequency rate for each word. Some stop words, as well as task-related words such as "submit", "agree", and "screen", are predominantly included in the highest mask frequency rate of words. Conversely, prepositions such as "at" and "in" exhibit a significantly lower masking rate compared to their token count, indicating their relatively important role in CPS recognition. This implies that removing stop words indiscriminately does not benefit dialogue act recognition, and it is important to learn how to mask because the importance of words varies by domain.

Effect of input length. We analyzed the performance of DP-FT5 for short and long inputs. We define short utterances as utterances with equal to or less than the median number of tokens (i.e., 4

Input	Size	FT5	DP-FT5	Diff.(%)	Avg. mask rate
Short	3375	46.6	64.0	17.4	17.7
Long	2595	59.0	62.8	3.8	10.3

Table 3: Average accuracy in short and long utterances with FT5 and DP-FT5 on the CPS dataset (median=4).

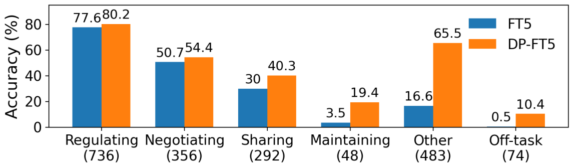


Figure 4: Average accuracy by class in the CPS dataset.

tokens) and long utterances for the remaining ones on the CPS data. Table 3 shows that DP-Masking benefits shorter utterances more than longer ones. Specifically, DP-FT5 improves by 17.4 percentage points over FT5 for shorter utterances and by 3.8 percentage points for longer ones. This could be because longer utterances, which show a lower masking rate of 10.3% compared to 17.7% for shorter ones, may not perform sufficient iterations to develop an effective masking strategy before training is halted to avoid overfitting, especially since both long and short utterances are trained together.

Accuracy by class. We studied how DP-Masking impacts performance improvement depending on each dialogue act class. Figure 4 displays the average accuracy by class for FT5 and DP-FT5. While the "other" category, characterized by unclear patterns, along with minority labels such as "maintaining", "sharing", and "off-task", show significant performance improvements, the majority labels such as "regulating" and "negotiating" exhibit only modest increasing trends. This implies that DP-Masking effectively handles noisy categories (i.e., "other") and enhances classification by masking specific tokens that would otherwise lead to misclassfication into other majority classes. This tendency is further demonstrated in the following qualitative analysis.

Qualitative analysis. We conducted a qualitative analysis to understand how DP-Masking benefits recognition. Below, we show two successful masking examples to illustrate how masking can mitigate bias due to token frequency when non-masking methods fail. Consider Successful Example 1. In our dataset, we found that the "regulating" class is twice as large as "negotiating" class. Also, although "agree" appears more frequently in "negotiating" (13%) than in "regulating" (8%), the absolute number of instances containing "agree" in "regulating" is 1.4 times bigger. This means the presence of "agree" can negatively impact "negotiating." However, DP-Masking successfully alleviates this issue by masking "agree" (in the third utterance). It also learns not to mask tokens like "anyone" that never appears in "regulating" class in the dataset. Hence, while non-masking FT5 misclassified this example as "regulating," DP-FT5 correctly classified it as "negotiating." We observed a similar phenomenon in Successful Example 2, where "do you" was over-represented in "regulating", like "agree" in the previous example. So DP-Masking learns to mask this phrase.

Successful Example 1.
speaker-A : how do you see the notes ?
speaker-B : Oh wait , I need ~~to~~ (inaudible) .
speaker-A : Does anyone want ~~to agree~~ with what I put ?
FT5: Regulating (Incorrect) | DP: Negotiating (Correct)
Successful Example 2.
~~speaker-C~~ : note.
speaker-D : ~~Do you~~ think that ?
~~speaker-C~~ : how ~~do you~~ see the notes ?
FT5: Regulating (Incorrect) | DP: Sharing (Correct)

As a counterexample to Successful Example 1, Failed Example 1 illustrates that masking "agree" in situations where the context and main utterance provide limited additional information can lead to ambiguity in the speaker's intention. In Failed Example 2, masking all the words in the main utterance fails to convey sufficient information.

Failed Example 1.
speaker-F : yes, I want it to (inaudible) .
speaker-G : I am very sorry .
speaker-F : Does everyone ~~agree~~ ?
FT5: Negotiating (Correct) | DP: Regulating (Incorrect)
Failed Example 2.
speaker-E : ~~[name]~~, you have to answer the question .
speaker-F : What question ? Oh, (.5) there's question .
speaker-E : ~~Y'all submit~~ .
FT5: Regulating (Correct) | DP: Sharing (Incorrect)

To summarize, imbalanced dialogue act datasets often suffer from overall token frequency significantly influencing the performance of dialogue act classifiers, rather than frequency within each class. Through the case studies above, we observed the potential of DP-Masking to mitigate such bias induced by token frequency.

7 Experiment: Public Benchmark

We also evaluate DP-Masking on three large-scale public datasets and analyze the impact of data size on the effectiveness of masking approaches.

Data. The Oasis dataset (Leech and Weisser, 2003) contains 1219 transcripts of live calls made to British Telecommunications and operator services with 42 dialog acts. We adopted the same data split as used in Chapuis et al. (2020). The TalkMoves dataset (Suresh et al., 2021) consists of 501 written transcripts of conversations between teachers and students in kindergarten through 12th grade math lessons with 6 talk move (dialogue act) labels. Switchboard Dialog Act (SwDA) (Godfrey and Holliman, 1993) is a telephone speech dataset with about 2,400 two-sided conversations among 543 speakers with 42 dialogue acts (Jurafsky et al., 1997). See Table 5 in Appendix C for the basic statistics of these datasets. None of the data have personal information.

Compared methods. We chose Flan-T5-small (FT5) as the underlying model because it is lightweight (low training and inference costs) and performs competitively on the CPS data. We compare the following masking approaches. (1) **FT5**: FT5 with no masking. (2) **R-FT5**: Random masking with FT5, following RoBERTa (Liu et al., 2019). (3) **DP-FT5**: Dual-process masking with FT5 (our proposed method). Hyperparameters are described in Appendix C.

Effect of Dataset Size. Figure 5 shows the classification performance of the above-mentioned methods on Oasis, TalkMoves, and SwDA with increasing sizes of the training and validation data while keeping the test set fixed. We averaged the results over 10 random seeds and report Mathews correlation coefficient (MCC) with standard deviation, following prior work (Suresh et al., 2021). We can see that DP-FT5 performs comparably or better than the two baselines. For smaller dataset sizes, DP-FT5 significantly outperforms the two baselines. As the data size increases, the benefit of both masking methods decreases. Interestingly, with sufficiently large data, the performance gap between DP-FT5 and the baselines almost disappears, suggesting that the strategically learned masking strategy through DP-Masking is effectively covered by exposure to more data. Thus, when data is limited, DP-masking can help achieve performance improvements that are equivalent to those obtained by additional data. Overall, across the three dialogue benchmarks with increasing amounts of data, DP-Masking outperforms random masking with an average normalized improvement of 3.2% as well as a no-masking strategy with 8.3% in MCC.

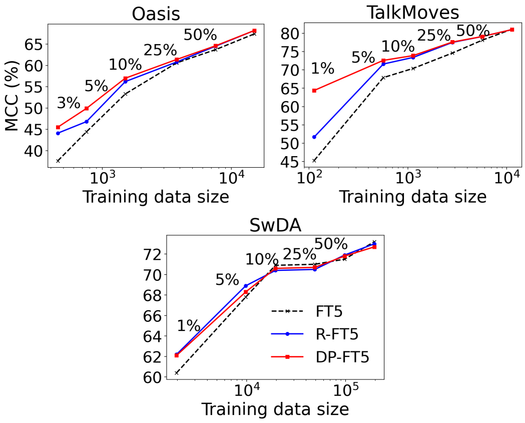


Figure 5: MCC scores relative to data size in three public benchmark datasets.

	Method (data size)	MCC (5%)	MCC (100%)	Acc (100%)
Oasis	FT5	44.4 (2.5)	67.4 (0.3)	70.5 (0.2)
	R-FT5	$*$46.8 (1.4)	$*$**68.2 (0.3)**	$*$67.5 (0.3)
	DP-FT5	§$*$**49.9 (0.8)**	$*$**68.2 (0.5)**	§**70.8 (0.4)**
	$HT(\theta^u_{MLM})$ (Chapuis et al., 2020)		-	69.4 (-.)
TalkMoves	FT5	67.9 (1.8)	**81.1 (0.2)**	90.3 (0.1)
	R-FT5	$*$71.6 (0.3)	81.0 (0.1)	90.2 (0.1)
	DP-FT5	§$*$**72.6 (1.7)**	81.0 (0.3)	90.2 (0.2)
	RoBERTa-Base (Suresh et al., 2022)		78.1 (-.)	-
	RoBERTa-Large (Suresh et al., 2021)		77.8 (-.)	-
SwDA	FT5	68.1 (0.6)	**73.2 (0.1)**	78.2 (0.1)
	R-FT5	$*$71.7 (0.4)	73.0 (0.1)	78.1 (0.1)
	DP-FT5	§$*$**72.1 (1.4)**	$*$72.7 (0.3)	77.7 (0.2)
	BERT+CC-FT (Noble and Maraev, 2021)		-	77.4 (-.)
	$HT(\theta^u_{MLM})$ (Chapuis et al., 2020)		-	79.3 (-.)
	Seq2Seq$_{\text{BEST}}$(Colombo et al., 2020)		-	**85.0 (-.)**

Table 4: Public benchmarks: dialogue act recognition performance on the 5% and the full data size. $*$ and § indicate statistical significance of the corresponding model compared to the baseline FT5 and R-FT5, respectively, using a Wilcoxon ranksum test ($p < 0.05$). The highest score in each column for each dataset is marked in bold.

Comparison with Prior Work. Table 4 presents the comparison with prior work on these datasets. On the 5% datasets, DP-FT5 significantly outperforms R-FT5, which, in turn, outperforms FT5 (Table 4). However, with the full data, the two masking approaches only surpass FT5 on the Oasis dataset, showing no improvement on the other datasets. Compared to prior work, DP-FT5 sets a new SOTA with 70.8% accuracy on the full Oasis dataset, surpassing the previous best of 69.4% (Chapuis et al., 2020). On TalkMoves, all three FT5-based methods achieve 81% MCC, surpassing the previous best of 78.1% (Suresh et al., 2021). On SwDA, all three methods perform similarly, with DP-FT5 offering no significant advantage over FT5 or prior methods. This suggests that DP-Masking is particularly beneficial for smaller datasets like Oasis (15K) and remains competitive on larger ones like TalkMoves (174K) and SwDA (199K).

In summary, DP-Masking excels on smaller datasets by focusing on essential structures and remains competitive on larger datasets. This approach improves dialogue act recognition and provides a denoising effect comparable to that of larger datasets. Analyzing masked data offers deeper insights into dialogue act recognition that traditional deep learning models may overlook, helping refine dialogue act taxonomies, improve annotation guidelines, and advance ongoing research. Overall, it provides a valuable foundation for investigating the linguistic characteristics of dialogue acts.

8 Conclusion

We have introduced the Dual-Process Masking (DP-Masking) framework for dialogue act recognition. DP-Masking consists of a reflective learning stage for extracting successful masks and an intuitive learning stage for encoding refined information into long-term memory. The results, validated across collaborative problem-solving dialogue data and three public dialogue benchmarks, reveal three key findings: (1) DP-Masking achieved improvements over competitive baseline models in challenging scenarios with limited data. (2) It demonstrated robust performance on public benchmarks, establishing a new SOTA for a benchmark task. (3) It enhanced interpretability by assessing the contributions of masked tokens to dialogue act recognition.

Future work can explore incorporating various reflection techniques to identify performance improvement points, integrating feedback on this work's findings into the learning process, and investigating whether DP-Masking enhances performance in other natural language processing tasks such as question answering, summarization, and document classification. The implementation of our work is available at *https://github.com/ykim32/DPMasking*.

Limitations

A limitation associated with DP-Masking is that it uses a pretrained generative language model as an underlying model to learn both a masker and a classifier, which requires a system with a capable GPU. Thus, while DP-Masking can theoretically be applied to any generative language model, it is most practical for smaller language models due to limited computing resources. Another limitation is that DP-Masking does not consider correlations among mask tokens when training a masker; it accumulates successful masks in order based on random initialization, potentially missing a better set of mask tokens for each utterance. This can be explored in future research on developing random token selection that considers correlations or collocations among tokens. Additionally, our datasets and all experiments are only focused on English dialogues. However, theoretically, it is expected to be applicable to other languages as well, which is a promising direction for future research.

Ethical Consideration

We do not directly observe ethical issues stemming from the technology introduced in this study. Since speaker information in the human dialogue data targeted by this technology is anonymized and not used, biases arising from such speaker information are not present. However, latent biases may still exist due to the less prioritized treatment of language habits and content of minority participants in model training. Additionally, this study is the first research on masking approaches for dialogue act recognition, and analysis related to minorities based on factors such as speaker's gender, race, or ethnicity is lacking. Addressing this issue in future research is important; however, it is worth noting that many dialogue datasets, including those used in this study, do no contain speaker information.

We respected the copyright and licensing terms of the existing models explored in this work. BERT, FT5, SGNN are under Apache License, CASA is under MIT License, RoBERTa is under GNU General Public License, and Llama 2 & 3 are under Meta's Community License Agreement.

Acknowledgments

This work is supported by the National Science Foundation under award DRL-2112635. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

A. Ahmadvand, J. Ingyu Choi, and E. A. Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd International ACM SIGIR Conference*, pages 1273–1276.

J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, volume 1, pages I/1061–I/1064 Vol. 1.

C. Bothe, C. Weber, S. Magg, and S. Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1952–1957, Miyazaki, Japan.

E. Chapuis, P. Colombo, M. Manica, M. Labeau, and C. Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2636–2648.

H. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

P. Colombo, E. Chapuis, M. Manica, E. Vignon, G. Varni, and C. Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

J. Devlin, M-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 1:4171–4186.

J. St BT. Evans. 2003. In two minds: Dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Z. Fan, Y. Gong, D. Liu, Z. Wei, S. Wang, J. Jiao, N.Duan, R. Zhang, and X. Huang. 2021. Mask attention networks: Rethinking and strengthen transformer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1692–1701.

J. J. Godfrey and E. Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium.

A. Goyal and Y. Bengio. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068.

G. Gronchi and F. Giovannelli. 2018. Dual process theory of thought and default mode network: A possible neural foundation of fast tinking. *Frontiers in Psychology*, 9(1237).

D. Hong, C. Feng, X. Zou, C. Hmelo-Silver, K. Glazewski, T. Wang, B. Mott, and J. Lester. 2024. Examining coordinated computer-based fixed and adaptive scaffolds in collaborative problem-solving game environments. In *Proceedings of the Seventeenth International Conference on Computer-Supported Collaborative Learning*, pages 43–50.

Y. Hu, C-H. Lee, T. Xie, T. Yu, N. A. Smith, and M. Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2627–2643.

D. Jurafsky, L. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallow discourse-function annotation: Coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

S. B. Kaufman. 2011. Intelligence and the cognitive unconscious. *The Cambridge Handbook of Intelligence*, pages 442–467.

S. Keizer, R. op den Akker, and A. Nijholt. 2002. Dialogue act recognition with Bayesian networks for dutch dialogues. In *SIGdial Workshop on Discourse and Dialogue*, pages 88–94.

S. Kim, L. Cavedon, and T. Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Z. Kozareva and S. Ravi. 2019. ProSeqo: Projection sequence networks for on-device text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3894–3903, Hong Kong, China. Association for Computational Linguistics.

H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3440–3447.

V. Kumaran, J. Rowe, B. Mott, S. Chaturvedi, and J. 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*, pages 10978–10992.

G. Leech and M. Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 441–446. Lancaster: Lancaster University.

Y. Levine, B. Lenz, O. Lieber, O. Abend, K. Leyton-Brown, M. Tennenholtz, and Y. Shoham. 2021. Pmi-masking: Principled masking of correlated spans. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

R. Li, C. Lin, M. Collinson, X. Li, and G. Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China.

J. Lin, W. Tan, L. Du, Wr. Buntine, D. Lang, D. Gašević, and G. Chen. 2024. Enhancing educational dialogue act classification with discourse context and sample informativeness. *IEEE Transactions on Learning Technologies*, 17:258–269.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Z. Liu, Z. Wang, Y. Lin, and H. Li. 2022. A neural-symbolic approach to natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2159–2172.

J. D. Marx and K. Cummings. 2007. Normalized change. *American Journal of Physics*, 75:87–91.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 23:276–282.

B. W. Mott, R. G. Taylor, S. Y. Lee, J. P. Rowe, A. Saleh, K. D. Glazewski, C. E. Hmelo-Silver, and J. C. Lester. 2019. Designing and developing interactive narratives for collaborative problem-based learning. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS Proceedings 12*, pages 86–100. Springer.

B. Noble and V. Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 166–172.

M. Nye, M. Tessler, J. Tenenbaum, and B. M. Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. 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.

V. Raheja and J. Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3727–3733, Minneapolis, Minnesota.

S. Ravi and Z. Kozareva. 2018. Self-governing neural networks for on-device short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 887–893, Brussels, Belgium. Association for Computational Linguistics.

T. Saha, A. Patra, S. Saha, and P. Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

R. Sun. 2015. Interpreting psychological notions: A dual-process computational theory. *Journal of Applied Research in Memory and Cognition*, 4(3):191–196.

D. Surendran and G.-A. Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *Proceedings of the 9th International Conference on Spoken Language Processing*, pages 1950–1953.

A. Suresh, J. Jacobs, C. Harty, M. Perkoff, J. H. Martin, and T. Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 4654–4662.

A. Suresh, J. Jacobs, V. Lai, C. Tan, W. Ward, J. H. Martin, and T. Sumner. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. In *the Spring AAAI 2021 Symposium on Artificial Intelligence for K-12 Education*.

M. Syed and S. C. Nelson. 2015. Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, 3:375–387.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

A. A. Von Davier, M. Zhu, and P. Kyllonen. 2017. *Innovative assessment of collaboration*. Springer International Publishing.

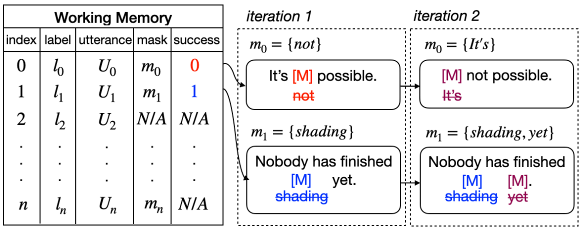
Y. Xu, E. Yao, C. Liu, Q. Liu, and M. Xu. 2023. A novel ensemble model with two-stage learning for joint dialog act recognition and sentiment classification. *Pattern Recognition Letters*, 165:77–83.

M. Zhao, T. Lin, F. Mi, M. Jaggi, and H. Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2226–2241.

A Method: Example

Figure 6 illustrates how successful masks are updated iteratively using the working memory W. Here, W is the same size as the entire training data U^{tr} and allows access to specific instances through indexing. Throughout each iteration, U^{tr} undergoes random partitioning into reflective training data U^{r_tr} and reflective validation data U^{r_v} to support the learning. In this process, M_R predicts masks for given utterances U_i, and then C_R classifies the masked utterances $M(U_i)$. W updates both the predicted masks and the classification results, using the "success" column in Working Memory, based on the current validation set (0: unsuccessful, 1: successful). In the subsequent iteration, correctly classified $M(U_i)$ acts as a base input, combined with new random masks to generate masked utterances featuring a heightened masking rate. Conversely, the mask tokens from incorrectly classified instances are discarded, and new random masks are introduced. In the example of Figure 6, during the first instance of iteration 1, "not" was randomly masked, but the dialogue act of that instance was misclassified, resulting in its success value of 0, and "not" was discarded. Then, a new mask "It's" is explored in iteration 2. On the other hand, during the second instance of iteration 1, "shading" was randomly masked, and the dialogue act of this instance was correctly classified. Thus, "shading" was recorded as a successful mask (success=1) and used for the next iteration. As a result, in iteration 2, the second instance has a successful mask token "shading" and a new exploring mask token "yet."

Through this iterative process, each instance is assigned a different masking rate depending on the significance of its tokens. Unlike our DP-Masking approach, random masking (i.e., single-process masking) executes random masking every iteration, relying solely on intuitive learning with no reflection on the results of the current masking strategy.



Working Memory				
index	label	utterance	mask	success
0	l_0	U_0	m_0	0
1	l_1	U_1	m_1	1
2	l_2	U_2	N/A	N/A
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	l_n	U_n	m_n	N/A

Figure 6: Example of how the working memory works to learn successful masks.

B Computational Cost and System

When using a pretrained language model as an underlying model for dialogue act classification, the computational cost of the baseline approach with no masking is $O(PNE)$ where P is the number of trainable parameters in the language model, N is the number of training instances, and E is the number of training epochs. The computational cost of random masking and our proposed DP-Masking approach are $O(PNEe_{max})$ due to the iterative nature of the algorithm, with a max number of training iterations e_{max}. Because DP-Masking involves training four components (M_R, C_R, M_I, C_I), its computational cost is a constant multiple of the random masking model.

We used multiple systems equipped with an NVIDIA GeForce RTX 2080 GPU with 8GB memory and an AMD EPYC 7302P CPU with 16 cores for experiments, except the Llama models, trained on a GPU, NVIDIA A10 or A30. Training times vary based on data size and hyperparameters. On average, DP-FT5 is about 7 times slower to train than FT5, while R-FT5 is 2.5 times slower. For example, processing 5% of the TalkMoves training data takes 20 minutes with DP-FT5, 7 minutes with R-FT5, and 3 minutes with FT5 on our system (RTX 2080 GPU). Although DP-Masking has a longer training time, it enhances performance in dialogue act recognition and provides insights into the tokens contributing to classification tasks, aiding the understanding of specific dialogue acts.

C Hyperparameters

We applied early stopping for training models, with a patience of 3 for all models except Llama, where the patience was set to 1. The best hyperparameter is highlighted in bold.

Collaborative problem-solving data For BERT models, the BERT-base-uncased models were optimized with AdamW using a learning rate of 3e-5. BERT uses a two layer feedforward network head for classification with a hidden layer size 512 (search space: [128, 256, **512**]). BERT-LSTM uses an LSTM with 512 neurons (fixed BERT output size // 2) and linear classification head. BERT-BiLSTM uses a similar architecture to BERT-LSTM, except the linear layer is increased to accommodate the concatenated bidirectional outputs.

For prior work, in RoBERTa, we searched the hidden layer of forward networks with [128, 256,

Dataset	(# classes)	CPS (6)	Oasis (42)	TalkMoves (7)	SwDA (42)
Number of instances	Total size	1,990	15,065	174,400	199,736
	Training size	1,492	13,589	151,184	195,658
	Test size	498	1,478	23,216	4,078
	Avg. size per class (SD)	332 (259)	377 (604)	24.9K (38.6K)	4.7K (13.2K)
Tokens	Avg. text length [min, max]	5 [1, 43]	9 [1, 420]	16 [1, 123]	7 [1, 107]
Hyperparameter	batch, learning rate	4, 0.0005	8, 0.0004	16, 0.0002	16, 0.0003

Table 5: Statistics of the classification benchmarks.

512], learning rate=[1, 3, **5**]$\times 10^{-3}$, and [8, 16, **32**] of batch size. For CASA-RoBERTa-base, we searched hidden = [**768**, 1024], batch = [**4**, 8, 16], learning rate=[1, 3, **5**]$\times 10^{-3}$ with 4 of number of workers. For SGNN, we searched d = [128, 160, **192**, 256], T = [60, **80**, 100], hidden = [128, **256**, 512], and batch = [64, **100**, 128]. See Kozareva and Ravi (2019) for details of the hyperparameters.

For the FT5-based models, we searched learning rate = [1, 3, **5**]$\times 10^{-3}$ and batch size = [**4**, 8, 16]. Then, we applied the best hyperparameters from the baseline FT5 to R-FT5 and DP-FT5. For the masking approaches (R-FT5 and DP-FT5), we explored the masking rate per input text, $p = [0.05, 0.1, 0.15, 0.2]$ using the training data. R-FT5 has a 5% masking rate and DP-FT5 has a 15% masking rate as the best on the CPS data. For Llama-2-7B and Llama-3-8B, we applied 8-bit quantization using QLoRA (r=8, alpha=32, dropout=0.1) with the same batch size, learning rate, and masking rate as the FT5 models.

Public benchmarks We set the max token number to 64 for each input in the dialogue datasets. In the base FT5 models, we explored learning rate = [1, 3, 5]$\times 10^{-3}$ and batch size = [4, 8, 16]. Then, we applied the best hyperparameters from the baseline FT5 to R-FT5 and DP-FT5 for each dataset. For the masking approaches (R-FT5 and DP-FT5), we explored the masking rate per input text, $p = [0.05, 0.1, \mathbf{0.15}, 0.2]$ using the training data. Each model was trained with early stopping (patience=3). The best hyperparameters are reported in Table 5.

D Metrics

Normalized improvement (Marx and Cummings, 2007) is defined as:

$$\begin{cases} (x-b)/b, & \text{for } b > x \\ (x-b)/(1-b), & \text{otherwise} \end{cases} \tag{1}$$

where x is a target value and b is a baseline value.

E Additional Results

This section includes additional results for the collaborative problem-solving (CPS) data. Figure 7 is a graph corresponding to the results of Table 1, intended for better visual comparison of models.

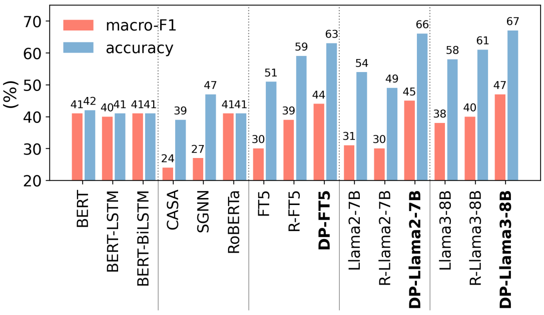


Figure 7: Classification performance on the dialogue acts for collaborative problem-solving data.

Figure 8 shows the mask distribution for each class by part of speech (POS) in the CPS and TalkMoves dataset. We observe that in the CPS data all the classes except "other" have similar POS mask distribution with a different scale where adjectives (ADJ), determiners and articles (DET), and pronouns (PRON) have the highest masking rates. On the other hand, in "other", nouns (NOUN) has the highest masking rate, followed by ADJ and PRON. This implies that many noun words mentioned in "other" are irrelevant to dialogue act classification.

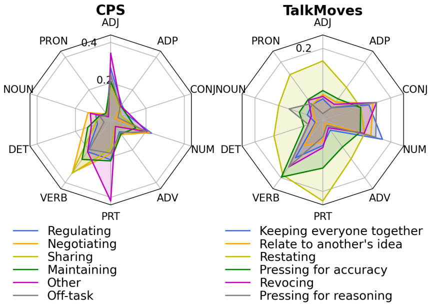


Figure 8: Mask distribution for each class by part of speech (POS) in the CPS and TalkMoves dataset.