Enhancing the Automatic Classification of Metadiscourse in Low-Proficiency Learners' Spoken and Written English Texts Using XLNet

Wenwen Guan[1] **Marijn Alta**[2] **Jelke Bloem**[2]
[1] Amsterdam Center for Language and Communication, University of Amsterdam
[2] Institute for Logic, Language and Computation, University of Amsterdam
w.guan@uva.nl, marijnalta@gmail.com, j.bloem@uva.nl

Abstract

This study aims to enhance the automatic identification and classification of metadiscourse markers in English texts, evaluating various large language models for the purpose. Metadiscourse is a commonly used rhetorical strategy in both written and spoken language to guide addressees through discourse. Due to its linguistic complexity and dependency on the context, automated metadiscourse classification is challenging. With a hypothesis that LLMs may handle complicated tasks more effectively than supervised machine learning approaches, we tune and evaluate seven encoder language models on the task using a dataset totalling 575,541 tokens and annotated with 24 labels. The results show a clear improvement over supervised machine learning approaches as well as an untuned Llama3.3-70B-Instruct baseline, with XLNet-large achieving an accuracy and F1-score of 0.91 and 0.93, respectively. However, four less frequent categories record F-scores below 0.5, highlighting the need for more balanced data representation.

1 Introduction

Metadiscourse (MD) is an essential rhetorical strategy in both speaking and writing that realizes two of the metafunctions of language proposed by Halliday (1994): the textual and interpersonal functions. MD that mainly has a textual function is used to form a cohesive and coherent text (Kopple, 1985). The textual dimension comprises transitions (e.g., *but*, *and*, *because*), frame markers (*firstly*, *in conclusion*, *the next point is ...*), code glosses (e.g., *in other words*, *namely*, *for example*), and so on. Textual MD markers often have fixed forms and consistent meanings, hence they pose relatively few challenges in automatic classification. Conversely, MD that is primarily interpersonal shows different features. Addressers use interpersonal MD to comment on the propositions and to involve the addressees in their discourse. Examples include but are not limited to hedges (e.g., *may*, *probably*, *I'm not sure ...*), boosters (e.g., *certainly*, *must*, *I believe*), and addressing the addressees (e.g., *You may end up thinking that ...*, *You may ask...*, *Can you hear me?*). This dimension is linguistically more complex as it involves multiple syntactic classes and has fuzzy span boundaries. The complexity undoubtedly leads to difficulty in automatic classification. Previous research using supervised methods reveals the performance gap between the two broad dimensions (dos Santos Correia, 2018; Alharbi, 2016). Classification of textual MD has yielded satisfactory accuracy but classification of interpersonal MD is lacking.

Automatic MD classification has barely been studied. We are only aware of the two aforementioned SVM-based studies, where transformer-based Large Language Models (LLMs) have not been used. As LLMs encode a broader range of semantic, syntactic and contextual information due to the more complex architecture that is pretrained on various linguistic resources, this study aims to improve the state of the art in automatic MD classification with a transformer-based method.[1]

The raw data used in this paper were sampled from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023). They effectively represent the diversity of natural language by containing both spoken and written English, as well as data from native speakers and learners with moderate proficiency in English. In linguistics, MD is often observed in non-native speakers' language. It has been found that learners may use more MD markers compared to native speakers because they intentionally put more effort into linguistic and meta-linguistic matters (Ädel, 2006). However, the majority of existing work about MD has focused on proficient speakers. This

[1] https://github.com/W-Guan/Automatic-MD-annotation-with-XLNet

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 27–41

study will pioneer research on MD used by learners with less advanced English proficiency.

Given that MD markers are captured in spans of varied lengths and are sensitive to context, we include a range of models in our study that are partly motivated by having a span-related training objective (SpanBERT: Joshi et al., 2020; ERNIE 2.0: Sun et al., 2020; XLNet: Yang et al., 2019), being state-of-the-art (ModernBERT: Warner et al., 2025), being a baseline for comparison (BERT: Devlin et al., 2019) and being common alternative choices (RoBERTa: Liu et al., 2019; ELECTRA: Clark et al., 2020). We also include Llama 3.3 70B-Instruct (Grattafiori et al., 2024) as an untuned decoder LLM baseline.

We find that the tuned encoder models show good performance for this task, with XLNet having the overall best accuracy and weighted F1-score, and ERNIE achieving a higher macro-F1 score, meaning better performance on low-frequency categories. Llama 3.3 70B-Instruct did not achieve adequate performance, even in a few-shot setting with per-category examples. Our models also outperform previous work, though this work used different MD categorization schemes and different kinds of text corpora.

2 Background

This section extends the concept of MD to its classification. Classification first refers to a theoretical taxonomy, which can be used when labelling the raw data. Two challenges have been identified in the theoretical classification and are anticipated to lead to difficulties in automatic classification.

2.1 Metadiscourse taxonomies

We introduced two main broad categories of MD. In linguistic practice, MD is classified into many categories but there is no uniform taxonomy. Along with the development of relevant research, scholars in the field came up with taxonomies from varied perspectives (Kopple, 1985; Crismore et al., 1993; Mauranen, 1993; Milne, 2003; Hyland, 2005; Ädel, 2006, 2010). Among them, Hyland's (2005) taxonomy is the most commonly used one. It provides a list of discovered MD markers for English based on corpus data. However, we had concerns about adopting the taxonomy in this study. Above all, it might not be sufficient because it is built on the observations of formal written language, including textbooks, students' writing, research articles, etc. Thus, it may not include MD markers typical of spoken language, nor incorporate mistakes such as grammatical errors and misspellings often found in learners' casual communication.

Ädel's (2010) taxonomy is representative of MD in spoken language. More importantly, the identification of the categories in this taxonomy relies primarily on the functions of MD in the discourse. For example, how the topic is introduced, developed, and concluded. Nevertheless, this taxonomy also has a limitation that does not fully meet the research purposes of this study. Ädel's (2010) taxonomy requires high explicitness of MD markers. If a text span does not contain deictic words that refer to the discourse or interlocutors, it will not be counted as MD in this taxonomy but may still be MD in other taxonomies. Therefore, we use Ädel's (2010) taxonomy as the basis of our annotation scheme but some categories from Hyland's (2005) taxonomy are added and adjusted. The final taxonomy we use is shown in Table 1. The dimensions of "Metalinguistic comments" and "Discourse organization" correspond to the textual MD, while the dimension "Writer-reader/speaker-listener interaction" aligns with the interpersonal MD.

The task of metadiscourse classification should not be confused with some related tasks that have been addressed in NLP. The task of dialogue act classification aims to label all of a dialogue according to their communicative function, such as 'request'. This can also include metadiscourse acts, but standard dialogue act classification schemes have been criticized for their unsystematic annotation of metadiscourse acts (Verdonik, 2023). Dialogue act classification is a NLP task where tuned encoder models outperform autoregressive decoder LLMs, which sometimes fail to beat rule-based baselines (Qamar et al., 2025), though one study shows limited success using ChatGPT in multiparty boardgame dialogue with a four-class scheme (Martinenghi et al., 2024).

Our task is also related to epistemic stance detection, which identifies statements that mark the writer's attitude towards the factuality of reliability of propositions (e.g. *I think that...*). Epistemic stance can be expressed through MD, but not all expressions of epistemic stance are MD markers. Epistemic stance detection also concerns propositions about others' stances, which fall outside the scope of MD. Several categories from our taxomony pertain to epistemic stances, specifically *Epistemic attitudes* (EPA), *Hedges* (HDG), *Boost-*

ers (BST) and *Speech act labels* (SAL). Eguchi and Kyle (2023) perform stance detection on an annotated corpus of student-written assignments as a span classification task, using the spaCy SpanCategorizer as a baseline and achieving best results with a RoBERTa-LSTM model, with results comparable to human inter-annotator agreement.

2.2 Challenges in MD classification

Two features of MD pose challenges in MD classification by human annotators and predictably also in automatic classification. Firstly, MD is highly context-sensitive. For instance, "so" functions as a MD marker when it indicates a causal relation between two clauses (Example 1.1). However, it is not a MD marker when it refers to a way something was described (Example 1.2). The ambiguity makes the identification of MD from propositional contents challenging.

Example 1.

- *1: All people in the restaurant would be affected by smoking* ***so*** *it should be banned.*
- *2: I don't think* ***so****.*

Secondly, a MD candidate may belong to more than one category. In Example 2, 'I fully agree that' is a MD marker to show the speaker's attitude. Within it, the MD marker 'fully' is a booster.

Example 2.

I fully agree that *smoking should be banned in restaurants.*

To date, the annotation of MD still heavily relies on manual annotation. Research on automatic classification remains highly limited. Two relatively in-depth studies have been conducted, focusing on MD classification in academic lectures (Alharbi, 2016) and TED talk transcripts (dos Santos Correia, 2018), respectively. dos Santos Correia (2018) used Support Vector Machines (SVMs) and Conditional Random Fields (CRFs). His study classified 10 categories and F1-scores for nine categories are below 0.6. Alharbi (2016) also used SVMs for primary exploration and then improved MD classification using Continuous Bag-of-Words (CBOW) and Convolutional Neural Networks (CNNs). In his study, three out of 19 categories got F1-scores higher than 0.8. Nonetheless, the model's predictions would not be reliable enough for further linguistic research. The biggest challenge of automatic classification lies in interpersonal MD because its syntactic and semantic features are more inexplicit and flexible compared to textual MD. For example, the *Exemplifying* (EXP) category achieved an accuracy of over 0.8 in both studies, while *Anticipating the audience's response* (AAR) got only 0.3 in Alharbi's (2016) work and even lower in dos Santos Correia's (2018) findings. We hypothesize that LLMs outperform these supervised methods.

Chan et al. (2024) also address metadiscourse using transformer models in the context of automated essay scoring. While they use a modified version of Hyland's (2005) classification scheme for manual annotation, they only perform a token-level identification task with a binary classification scheme. They find little difference between the performance of BERT, DistilBERT and RoBERTa on the task, and focus on under/oversampling techniques and different classification algorithms such as multi-layered perceptrons and AdaBoost. While useful for automated essay scoring, this identification task has limited utility for linguists who wish to construct a metadiscourse-annotated corpus.

2.3 Selected LLMs

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019) introduced bidirectional encoding by masked language modeling (MLM) and next sentence prediction (NSP). BERT and its variants, including ModernBERT (Warner et al., 2025), SpanBERT (Joshi et al., 2020), and RoBERTa (Liu et al., 2019), consist of stacked encoders that are trained on unlabeled data to encode contextualized language representations. With a token classification head, they have been used for a wide range of token and span labeling tasks, such as named entity recognition and part-of-speech tagging. Metadiscourse classification is part of the same family of tasks.

SpanBERT is trained with a span-boundary training objective. This encourages the model to represent the relationships between tokens within a span. It predicts the entire span of tokens instead of individual tokens. This is useful for tasks that require representations of text chunks, such as our task of MD identification and classification. Although RoBERTa and ModernBERT are not pretrained with span-specific boundaries, the optimization of model architecture and training such as dynamic masking and larger datasets makes them outperform the traditional BERT model. It remains unknown if this general optimization would lead to a better performance than span-specific pretraining.

Dimension	Category	Label	Examples
Metalinguistic comments	Repairing	RPR	*I'm sorry...*
	Reformulating	RFL	*to put it differently...*
	Commenting	CMT	*... is a difficult question.*
	Clarifying	CLF	*I don't mean to say*
	Exemplifying	EXP	*for example*
Discourse organization	Managing topics	MNT	*I will focus on...*
	Organizing statements	ORS	*and; but; so*
	Providing evidentials	PED	*according to*
	Enumerating	ENM	*first; at last*
	Endophoric marking	EDP	*As we can see in Chapter III, ...*
	Previewing	PRV	*We will discuss...*
	Reviewing	RVW	*As I said last time, ...*
Writer-reader /Speaker-listener interaction	Epistemic attitudes	EPA	*I agree that...*
	Hedges	HDG	*perhaps; might*
	Boosters	BST	*definitely; should*
	Speech act labels	SAL	*I argue that...*
	Managing comprehension	MNC	*You know what I mean.*
	Managing channel/audience discipline	MCD	*Can you hear me?*
	Anticipating the audience's response	AAR	*You may ask...*
	Managing the message	MNM	*What I want to emphasize is...*
	Imagining scenarios	IMS	*Suppose you're giving a speech...*

Table 1: List of MD categories, labels, and examples

We selected two other models with pretraining objectives that have particular relevance to our task. ERNIE's continual multi-task learning component includes two objectives relevant for our task. Firstly, there is a knowledge masking task, which requires the model to learn to predict masked spans and masked named entities rather than just tokens. Secondly, there is a discourse relation task, which relies on Sileo et al.'s (2019) discourse marker dataset to have the model predict the rhetorical relation between sentences during pretraining. This is related to metadiscourse, as discourse markers are mostly textual metadiscourse markers.

XLNet (Yang et al., 2019) is an autoregressive pretraining method of which the objectives include span-based prediction, where consecutive spans of up to five tokens are predicted rather than just single tokens. As textual MD markers are often spans, this may facilitate MD classification. However, MD spans go beyond the length of five tokens as well.

Lastly, ELECTRA (Clark et al., 2020) provides another alternative to masked language modelling pretraining by basing pretraining on a replaced token detection task. This approach is shown to outperform SpanBERT and perform similarly to XLNet on the somewhat related SQuAD benchmark (Rajpurkar et al., 2016), where models select spans that answer questions. Therefore, we also expect good performance on our MD classification task.

In recent years, autoregressive decoder-only generative LLMs have shown impressive generalization performance on a range of NLP tasks with few-shot prompting, even to novel tasks and domains without fine-tuning. However, they show poor results in these settings on text classification tasks (Bucher and Martini, 2024), span labeling tasks such as named entity recognition (Keraghel et al., 2024) and other tasks related to ours such as dialogue act classification (Qamar et al., 2025) and implicit discourse relation annotation (Yung et al., 2024). Nevertheless, we include Llama-3.3-70B-Instruct (Grattafiori et al., 2024) as a decoder LLM baseline. We also include a SpaCy baseline.

3 Methods and data

Our data are English learners' speaking and writing extracted from the ICNALE corpus. The selection of this corpus was initially motivated by an interest in its potential for subsequent qualitative analysis, specifically in examining learners' use of MD. The corpus has four modules, namely spoken monologues (SM), spoken dialogues (SD), writing (WR),

Module	Texts	Tokens	Avg. Tokens
SM	999	128,989	129.12
SD	748	162,759	217.59
WR	1,100	254,064	230.97
EE	130	29,729	228.68

Table 2: The descriptive statistics of the dataset

and edited writing (EE). The extracted data concern the four modules and the groups whose first language is Chinese (Mandarin or Cantonese) and English, which includes CHN (Chinese mainland), HKG (Hong Kong), TWN (Taiwan), SIN (Singapore), and ENS (English native speakers). Random sampling was used to select half of the data for manual annotation. After sampling, the data comprises 2,977 texts totaling 575,541 tokens excluding punctuation. Table 2 shows detailed statistics.

3.1 Annotation

The annotation scheme consists of the 21 MD labels from the taxonomy in Table 1. Three additional labels pertaining to linguistic errors made by the writers and ambiguities, including Grammatical errors (ERR), Misuse (MIS), and Uncertainty (UCT), were also annotated but have been fully addressed in the gold standard corpus. Thus, they are excluded from the present experiments. Manual annotation was conducted using Prodigy (Honnibal et al., 2024), an annotation tool for creating training and evaluation data for machine learning models. There are two annotators who have degrees in linguistics-related subjects. They were trained in the definition and classification of MD, difficult examples, and the use of Prodigy. Their annotation quality was evaluated by inter-annotator agreement (IAA) using the Cohen's kappa coefficient (κ). This metric for pairwise agreement, which accounts for chance agreement, was computed per token rather than per span in order to allow for partial matching. Table 3 reports the IAA of the overall dataset, along with the label distribution which is visualized in Appendix 4. We observe strongly imbalanced class frequencies, which is a consequence of annotating natural language corpus data.

The macro average κ coefficient for all the MD categories is 0.79. This suggests that the majority of MD markers can be properly identified and classified. The disagreement is mainly attributed to the context-sensitive nature and fuzzy boundaries of MD. Taking 'I think' as an example, it is a

Label	κ	N-A1	N-A2
RPR	0.81	37	35
RFL	0.78	107	157
CMT	0.86	1,067	827
CLF	0.92	793	867
EXP	0.74	1,080	1,750
MNT	0.91	2,829	3,202
ORS	0.93	23,230	23,890
PED	0.88	2,077	2,488
ENM	0.95	4,887	5,076
EDP	0.79	152	173
PRV	0.94	788	801
RVW	0.89	783	866
EPA	0.88	15,039	14,791
HDG	0.85	9,330	9,547
BST	0.78	8,263	10,314
SAL	0.82	921	1,084
MNC	0.84	1,300	1,089
MCD	0.86	1,523	1,877
AAR	0.91	456	449
MNM	0.74	249	302
IMS	0.90	131	159
Macro Avg.	0.79		

Table 3: Pairwise Cohen's κ coefficients of inter-annotator agreement (IAA). N refers to the number of annotated tokens in spans of the specific label, whereas A1 and A2 are annotators.

Hedge (HDG) marker when it appears at the end of a clause, but it is marked with *Epistemic attitudes* (EPA) when it starts a clause due to its neutral tone. Furthermore, fuzzy boundaries are found among EPA markers, such as 'I agree (with)' and 'I agree (that)'. In this case, 'with' and 'that' should be included in the span. Label disagreements and inconsistent boundaries were resolved by discussion and the involvement of a third linguist.

We split the dataset into a 70/15/15 division for training, testing, and validation. We use splits with fixed random seeds for reproducibility and to ensure that every MD category is present in the validation and test set by re-splitting with a different fixed seed until this is the case.

3.2 Models

For the classification task, we use the base and large version of the aforementioned models: BERT, SpanBERT, ModernBERT, RoBERTa, ELECTRA, ERNIE and XLNet. We used cased models to facilitate the identification of sentence boundaries. We then tune these models on the task of MD classifi-

cation using the training portion of our corpus.

Specifically, we use these models to jointly perform the identification and classification tasks using token classifier heads tuned on the task. Predicting a span where no span was annotated is considered an incorrect prediction. In tuning, every token that is not covered by an annotated MD span in the training data, is considered an unlabeled (category '-') span. For our evaluation metrics we consider both weighted F1 (also called micro F1) and macro F1, due to strong class imbalance. Macro F1 weighs all metadiscourse categories equally, including those with only a few instances. With our imbalanced dataset, this metric emphasizes performance on small categories. Weighted F1 is weighted by the frequency of the category, emphasizing performance on larger categories.

For hyperparameter tuning, we use Bayesian optimization with HyperOpt, and an Asynchronous Successive Halving scheduler to increase the efficiency of the process. We tune the learning rate $\sim$ log-uniform$(10^{-5}, 50^{-5})$, weight decay $\sim U(10^{-3}, 10^{-1})$, training batch size $\sim U\{4, 32\}$ and warmup steps $\sim U\{4, 32\}$. We perform the tuning with 60 sample trials, 20 initial points and using the weighted F1-score as a metric. We run the models for 25 epochs. For evaluation, we use a batch size of 16. Best obtained hyperparameter combinations can be found in Appendix G.

For the Llama-3.3-70B-Instruct baseline, we adapted the widely used GPT-NER (Wang et al., 2025) sequence generation approach for named entity recognition to the task of metadiscourse classification. Details on our few-shot prompting approach with class label explanations are described in Appendix B. We use the default temperature hyperparameter of 0.6.

For the SpaCy baseline, included to represent supervised classifiers as used in previous work, we experimented with the SpanCategorizer (spaCy, 2024) pipeline. Integrated in Prodigy, it is convenient for corpus linguists who are not proficient in programming to perform automatic annotation. The span categorizer uses Tok2Vec embeddings as features with a vocabulary of 5000 (1000 prefix, 2500 suffix), going into a Maxout Window Encoder. We used a hidden layer size of 128, four encoding layers and a max span size of 22. It was trained with a 70-30 split of data for 20 epochs with the Adam optimizer, a learning rate of 0.001 with 0.01 weight decay and 10% dropout.

4 Results

For the SpaCy baseline, we observed the effects of the class imbalance inherent in our task. Only the five most frequent labels (ORS, ENM, EPA, HDG, BST) showed non-zero performance, while other labels were not predicted. The baseline's accuracy based solely on the non-zero values achieved 0.81 (computed per span), but drops to 0.19 when all categories are considered. This indicates that SpanCategorizer fails to generalize across all MD categories, and this result is inadequate for assisting linguists in semi-automatic MD classification. The complete results are presented in Appendix A.

The Llama-3.3-70B-Instruct baseline also performed poorly with an accuracy of 0.26, weighted F1-score of 0.41 and macro F1-score of 0.17 (Table 4). To confirm this result, we performed a follow-up experiment with the potentially more powerful GPT-4o model (Hurst et al., 2024) on 20 test documents and observed a weighted F1-score of 0.56 and a macro F1-score of 0.26, outperforming Llama-3.3-70B-Instruct but still trailing behind the fine-tuned encoder models. Further details on our decoder model baseline experiments can be found in Appendix B.

4.1 Model comparison

Table 4 shows the results of the base models on the metadiscourse classification task. In terms of accuracy and weighted F1-score, XLNet with its span-based prediction pretraining objective clearly outperforms the other models. However, its macro F1-score lags behind that of the other models, indicating that the model is relatively good at predicting common MD categories and relatively bad at predicting uncommon ones. In terms of macro F1, ERNIE-v2 shows the best performance.

In Table 5, results for the large versions of these models are shown. Results pattern similarly, with either no or very minor performance gains for most models. This indicates that the bottleneck for MD classification is in the classification head rather than in the base model, due to the relatively small amount of labeled data available.

Next, we examine the per-class performance for the best-performing XLNet-large model in Table 6. Unlike the spaCy and Llama-3.3 baselines, XLNet is able to classify all of the MD categories to some extent, even those with less than 50 labeled tokens in the test set. Nevertheless, we can observe performance issues due to class imbalance – some low

Model-base	Acc.	F1	MacroF1
BERT	0.871	0.901	0.751
SpanBERT	0.856	0.893	0.738
ModernBERT	0.858	0.897	0.752
RoBERTa	0.868	0.900	0.724
ELECTRA	0.863	0.898	0.739
ERNIE	0.870	0.904	**0.766**
XLNet	**0.905**	**0.922**	0.707
Llama-3.3-70B	0.257	0.409	0.168

Table 4: Evaluation results for different base models

Model-large	Acc.	F1	MacroF1
BERT	0.868	0.900	0.742
SpanBERT	0.857	0.894	0.743
ModernBERT	0.860	0.898	0.748
RoBERTa	0.869	0.901	0.741
Electra	0.846	0.890	0.754
ERNIE	0.866	0.900	**0.769**
XLNet	**0.915**	**0.930**	0.714

Table 5: Evaluation results for different large models

and mid-frequency categories exhibit F1-scores below 0.5. Figure 2 plots the relationship between the amount of instances in a category (frequency) versus the F1-score to visualize this pattern. We can observe a clear correlation between the two variables. All low-performing categories (F1<0.88) have 1000 tokens or less of support in the test set, reflecting the distribution in the training set.

This raises the question as to what causes these differences in performance in the lower frequency bracket. One potential explanation is lexical variability – categories that can be expressed by a larger range of words should be more difficult to classify.

4.2 Unique spans per category

An interesting property of MD categories is that some categories have far less variation than others. Textual MD markers are often grammaticalized and fixed in form, while interpersonal MD can be expressed in many ways, as discussed in the introduction. Therefore, we also examine the effect of MD variation on performance per category. Figure 3 plots each category in terms of their ratio of unique spans (variation), controlled for frequency, against the F1-score. Frequency is controlled by dividing the number of unique spans by the total amount of spans of that category, similar to how type/token ratio is computed. So, for example, the *Anticipating the audience's response* (AAR) category has a ratio of 1, meaning that every instance of it uses a unique sequence of tokens.

Label	P	R	F1	N
RPR	0.70	0.03	0.06	213
RFL	0.97	0.83	0.89	35
CMT	0.99	0.98	0.98	486
CLF	0.87	0.29	0.43	266
EXP	0.99	0.96	0.98	358
MNT	0.77	0.51	0.61	1,265
ORS	1.00	0.92	0.96	5,613
PED	0.98	0.86	0.92	623
ENM	0.92	0.92	0.92	696
EDP	1.00	0.67	0.80	15
PRV	0.68	0.33	0.44	234
RVW	0.98	0.97	0.97	289
EPA	0.99	0.99	0.99	17,130
HDG	0.99	0.93	0.96	4,358
BST	0.92	0.85	0.88	3,760
SAL	0.90	0.87	0.89	306
MNC	0.30	0.18	0.22	376
MCD	0.96	0.93	0.94	366
AAR	0.99	0.33	0.50	224
MNM	0.84	0.65	0.73	55
IMS	0.60	0.60	0.60	10
Acc.			0.91	36,678
Macro	0.83	0.67	0.71	36,678
Weighted	0.96	0.91	0.93	36,678

Table 6: Results per class for XLNet-large

We expect that the more varied categories are more difficult to classify, but in the top left corner we see that some of the most diverse categories also have high F1-scores, even when controlled for frequency. The *Reviewing* (RVW) and *Commenting* (CMT) categories show that even a highly variable categories can still get a high F1-score, while a category with somewhat lower variation, *Managing comprehension* (MNC) gets a lower F1-score while being similar in frequency to the aforementioned two categories. This indicates that the model has good generalization capabilities and does more than just remembering some common words for each category. We also note a pattern of more frequent categories, e.g., *Epistemic attitudes* (EPA) and *Boosters* (BST), having a lower variation ratio – this is tied to the frequency factor, as among more spans there are more likely to be repeated ones.

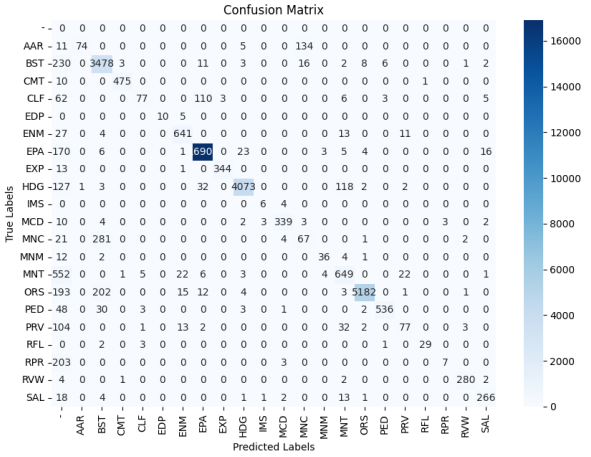

Figure 1: Confusion matrix for XLNet-large

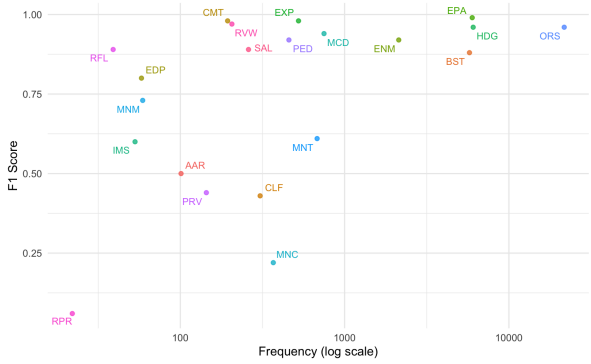

Figure 2: Relationship between label frequency and F1-score

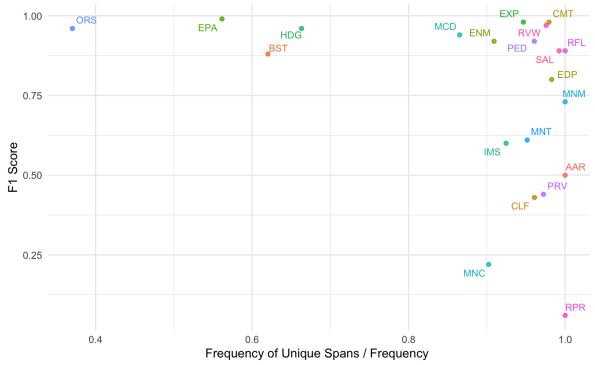

Figure 3: Relationship between the number of unique spans per category divided by frequency, and F1-score

4.3 Class ambiguity

Table 7 contains an overview of the most common misclassifications, highlighting elements from Figure 1. Such classification mixups indicate ambiguity between the categories, possibly because of similar MD spans. The overview includes misidentifications, where an annotated span is identified as not being MD — this is indicated with '-'. For each class, we show its frequency (N), the most common misclassification (Err1), what percentage of tokens of this class was misclassified in this way, and the second most common misclassification (Err2).

The first six categories are the relatively most commonly misclassified categories, and the correlation with category frequency is visible. For the least frequent class, *Repairing* (RPR), we see that almost all instances are misidentified. The bottom three categories are the three biggest categories. We can see that the most misclassified categories are misclassified as other uncommon categories, except *Clarifying* (CLF) which gets confused for *Epistemic attitudes* (EPA), the most common category. So, these confusions are more than a frequency effect.

Observed ambiguity instead appears to be caused by frequent overlapping of the two labels. They may co-occur in the same span a lot, as in our Example 2, where a *Boosters* (BST) occurs inside an *Epistemic attitudes* (EPA) indication. XLNet-large misclassifies a BST as EPA 11 times while classifying EPA as BST 6 times. Another situation caused by overlapped labels is the confusion between *Managing comprehension* (MNC) and BST. The misclassification particularly happens to the phrase 'As we all know', which is marked as both MNC and BST by human annotators.

Another potential cause of confusion is semantic ambiguity and similarity. For the ORS-BST confusion (*Organizing Statements-Boosters*), we observe that ORS spans that get misclassified as BST often contain words such as 'especially' and 'even', which are also used as boosters. To cite examples from the training data, the true label for 'even still' is BST, whereas that for 'even if' or 'even though' is ORS. The misclassification of *Anticipating the audience's response* (AAR) as MNC is due to the same reason. Looking at some of the AAR markers (e.g., *you may say*, *if you ask me*, *you might say that*) and the MNC markers (e.g., *as you know*, *as you can see*, *you mean*), we could find all the occurrences of the personal pronoun 'you' refer to the readers/listeners and functions as the subject in the sentence. This semantic similarity and use of 'you' by both categories leads to confusion. Similarly, the confusion between *Clarifying* (CLF) and *Epistemic attitudes* is caused by the common use of 'I'. These confusions are not observed in human annotators' performance.

Cat	N	Err	%	Err2
RPR	213	-	95.31%	MCD
MNC	376	BST	74.73%	-
AAR	224	MNC	59.82%	-
PRV	234	-	44.44%	MNT
MNT	1,265	-	43.64%	ENM
CLF	266	EPA	41.35%	-
HDG	4,358	-	2.91%	MNT
ORS	5,613	BST	3.60%	-
EPA	17,130	-	0.99%	HDG

Table 7: Two most common misclassifications and misidentifications of categories for the 6 most misclassified categories and the 3 largest ones, by XLNet-large.

5 Discussion

Overall, we observed that tuned encoder LLM-based approaches are able to perform metadiscourse identification and classification well, clearly outperforming the spaCy baseline (weighted F1-score of 0.93 vs 0.81 only for the most frequent categories) and the Llama-3.3-70B-Instruct decoder LLM baseline (weighted F1-score of 0.93 vs 0.41). Only the encoder models are able to classify all MD categories, including the more challenging interpersonal ones. The best performance was reached with XLNet-large, but performance gains from using large versions of models were minimal.

Direct comparison to previous SVM-based work (Alharbi, 2016; dos Santos Correia, 2018) is difficult as different MD categorization schemes were used, and F1-scores were only reported per category. Broadly, dos Santos Correia (2018) reports F1-scores below 0.6 in 9 of 10 categories, while XLNet only goes below 0.6 in 5 of 21 categories. Alharbi (2016) reports 3 of 19 categories with F1-scores over 0.8, while XLNet-Large achieves this for 13 of our 21 categories. As a challenging example, the *Anticipating the audience's response* (AAR) category got a F1-score of 0.3 in Alharbi's (2016) work, while XLNet gets 0.5.

Improved LLM-based metadiscourse classification has promising applications. First, we can use this method to expand our annotated corpus, which is still under development, to overcome the research limitations imposed by the small amount of available data. Moreover, automatic MD classification is beneficial to language acquisition research, particularly for second language acquisition. It can help with tasks related to (automated) language assessment and language teaching. It also has potential to be used in text analysis tools to provide language learners with concrete feedback on language coherence and interactionality. MD classification also provides an indication of the viability of the annotation of similar pragmatic and discursive properties of texts, in other words, linguistic items that are highly context-dependent and potentially challenging even to state-of-the-art NLP methods. Explicit MD representation may have other downstream potential for natural language processing tasks, such as in dialogue systems. It may also provide informative input features for related tasks such as stance detection, where stance can be expressed through metadiscourse, or dialogue act segmentation and classification, where some dialogue acts may be metadiscourse acts.

The practical applicability of our results is limited by poor performance on certain low-frequency classes. In future work, targeted supplemental annotation for the more challenging MD categories is the most promising next step. Another future research direction is improving generative decoder LLM performance on this task, such as through instruction tuning, prompt engineering or an agentive approach with domain-expert agents.

6 Conclusions

The model discussed in this paper achieved an accuracy and F1-score of 0.91 and 0.93 respectively on the task of metadiscourse identification and annotation, representing an improvement over related work, the SpaCy baseline and generative decoder LLM performance. Obtaining this performance with a fairly small annotated dataset shows that LLMs have potential to help speed up the annotation process even for fairly uncommon NLP tasks such as ours. The main open issue with this task is class imbalance. Annotating more conversations where these categories are present is the most promising approach to improve this.

7 Limitations

The main limitation of this project is the insufficient number of instances for the majority of categories. Annotating more conversations to make sure all categories have a significant number of instances is the most promising next step, as mentioned earlier. However, low frequency is inherent to some MD categories in natural language and small datasets with imbalanced category distributions are common in linguistic research. Alternative methods

should also be considered, for example, targeted annotating of only low-categories which would be more efficient.

A further limitation is that the study is based on a single dataset. Therefore, we cannot make claims about generalization to other learner populations or other discourse types. No other annotated MD datasets exist and it is costly to annotate them, though our results suggests that semi-automatic annotation can be performed after tuning on a relatively small dataset for future annotation of other corpora for MD.

The applicability of the method is limited by the fact that it requires tuning on MD annotation, a type of annotation that is not commonly available, especially in under-resourced languages. This limits the extent to which our method can be applied in diverse linguistic contexts and domain contexts. We were only able to demonstrate it for English. With our findings, semi-automatic annotation of more data can be explored, but only for languages where usable MD annotation already exists. Few-shot performance through in-context learning by a large generative LLM was insufficient to provide a viable alternative.

While MD spans can be embedded inside each other, the token classification approach adopted in this study can only assign one label per token. These cases are therefore not fully handled. The classifier can put a smaller span of category B inside a larger span of category A, but without the ability to assign multiple labels to one token, it is unspecified whether the category A span encompasses the category B span, or is actually two different category A spans, one before the category B span and one after.

We investigated the effect of span variation on classification performance, but our approach has some limitations. For example, if two instances of spans use the same words but in different order, or the same sequence of words but with one word omitted, they would count as different unique spans. Some sort of overlap or semantic similarity-based metric might give a better view of uniqueness of spans.

We discuss the application of our method to (semi)-automatic metadiscourse annotation. However, if it were to be used for this purpose, the annotations would be biased based on the mistakes that the classifier makes - categories that the classifier struggles with more, would be more poorly annotated. Annotators should pay particular attention to these underperforming categories.

References

Annelie Ädel. 2006. *Metadiscourse in L1 and L2 English*. John Benjamins.

Annelie Ädel. 2010. Just to give you kind of a map of where we are going: A taxonomy of metadiscourse in spoken and written academic English. *Nordic Journal of English Studies*, 9(2):69–97.

Ghada Alharbi. 2016. *Metadiscourse tagging in academic lectures*. Ph.D. thesis, University of Sheffield.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

Sathena Chan, Manoranjan Sathyamurthy, Chihiro Inoue, Michael Bax, Johnathan Jones, and John Oyekan. 2024. Integrating metadiscourse analysis with transformer-based models for enhancing construct representation and discourse competence assessment in L2 writing: A systemic multidisciplinary approach. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue):318–347.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Avon Crismore, Raija Markkanen, and Margrat S. Steffensen. 1993. Metadiscourse in persuasive writing: A study of texts written by american and finnish university students. *Written Communication*, 10(1):39–71.

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.

Rui Pedro dos Santos Correia. 2018. *Automatic Classification of Metadiscourse*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Masaki Eguchi and Kristopher Kyle. 2023. Span identification of epistemic stance-taking in academic written English. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

M.A.K. Halliday. 1994. *An Introduction to Functional Grammar*. Hodder Arnold.

Matthew Honnibal, Ines Montani, et al. 2024. Prodigy: An annotation tool for AI, Machine Learning & NLP. https://prodi.gy.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Ken Hyland. 2005. Metadiscourse: Exploring interaction in writing. *Continuum*.

Shin'ichiro Ishikawa. 2023. *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*.

William Vande Kopple. 1985. Some exploratory discourse on metadiscourse. *College Composition & Communication*, 36(1):82–93.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, Dimitri Ognibene, et al. 2024. LLMs of Catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multiparty dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing@ LREC-COLING 2024*, pages 107–118. ELRA and ICCL.

Anna Mauranen. 1993. Cultural differences in academic discourse–problems of a linguistic and cultural minority. *AFinLAn vuosikirja*, pages 157–174.

ED Milne. 2003. Metadiscourse revisited: a contrastive study of persuasive writing in professional discourse. regreso al metadiscurso: estudio contrastivo de la persuasión en el discurso profesional. *Estudios ingleses de la Universidad Complutense*, 11:29–52.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486.

spaCy. 2024. Span categorizer: Pipeline component. https://spacy.io/api/spancategorizer. Accessed: 2024-10-14.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Darinka Verdonik. 2023. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2):144–171.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

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

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165.

A SpaCy baseline

Label	Precision	Recall	F1-score
RPR	0.00	0.00	0.00
RFL	0.00	0.00	0.00
CMT	0.00	0.00	0.00
CLF	0.00	0.00	0.00
EXP	0.00	0.00	0.00
MNT	0.00	0.00	0.00
ORS	0.89	0.88	0.88
PED	0.00	0.00	0.00
ENM	0.96	0.44	0.60
EDP	0.00	0.00	0.00
PRV	0.00	0.00	0.00
RVW	0.00	0.00	0.00
EPA	0.91	0.90	0.90
HDG	0.87	0.73	0.79
BST	0.88	0.70	0.78
SAL	0.00	0.00	0.00
MNC	0.00	0.00	0.00
MCD	0.00	0.00	0.00
AAR	0.00	0.00	0.00
MNM	0.00	0.00	0.00
IMS	0.00	0.00	0.00
Acc.			0.81
Macro (Non-zero)			0.79

Table 8: Classification accuracy of the SpaCy baseline

B Decoder LLM baseline

For the Llama-3.3-70B-Instruct baseline, we adapted the popular GPT-NER (Wang et al., 2025) prompting approach from the Named Entity Recognition (NER) task to our task. The model is prompted to apply labels in this format: *No, @@I don't think#EPA## so*. In this example, EPA is the label, @@ the span start token and ## the span end token. We prompt for the label to be generated at the end of the span due to the unidirectional nature of generative decoder LLMs. In our prompt, we added a definition of metadiscourse, the persona phrase "You are an excellent linguist." from Wang et al. (2025), and short descriptions of all the categories with one example per category, which are the same as in Table 1. We used a three-shot setting, adding three randomly chosen full annotated documents from our training set (which was otherwise not used in the decoder LLM experiments). We also tried a zero-shot setting without the full annotated documents, but performance was very limited in these trials. The full three-shot prompt template can be seen in Appendix F.

In Table 9, we show results per class for Llama-3.3-70B-Instruct. The overall metrics are notably lower than those for the tuned encoder models. Four categories are never labeled correctly. As might be expected in a setting without fine-tuning, per class performance does not correlate as clearly with the class's frequency in our corpus — the rather frequent class of *Managing topics* (MNT) is never predicted by the model. The best classification performance is shown for classes that often consist of one or two words, such as *Enumerating* (ENM - *first*, *at last*), *Organizing statements* (ORS - *and*, *but*) and *Boosters* (BST - *definitely*, *should*). This is despite the fact that evaluation metrics are computed per token, and it suggesting difficulties in accurately marking boundaries of longer spans or in structure prediction more broadly.

As this result may be surprising to some readers, we performed an additional small-scale experiment with GPT-4o for verification. With a test set of 20 random documents (different from the example shots) and in the three-shot setting, GPT-4o achieved a weighted F1 score of 0.554 and a macro F1 score of 0.259, outperforming Llama-3.3-70B-Instruct but still underperforming compared to the tuned encoder models.

There are a few possible reasons for this poor performance. Firstly, our task is far less common than the NER task, so it is less likely to occur in training or instruction tuning data for these models. Secondly, our task has more categories than the NER task and the categories are domain-specific. They require specialized knowledge to interpret and are not frequently discussed outside of specialized literature. Even with the description and examples, the model does not seem to have an accurate representation of the categories.

Specifically, the model seems to experience interference from the more common tasks of discourse act labeling and discourse/conversation analysis. Even with the prompt stating that it's a metadiscourse task, the model often tries to perform a discourse analysis task and interprets our categories as if they are part of such a task. For example, it marks disfluencies as RPR (*Repairing*), while only discourse about disfluencies (metadiscourse) should be marked as such. An example of this from our GPT-4o evaluation is: "*In university, most – most @@students#RPR## first goal is to study*". This labeling is likely triggered by the repetition of *most*, but it is a repetition, not a repair (if we are do-

Label	P	R	F1	N
RPR	0.07	0.50	0.13	213
RFL	0.16	0.44	0.24	35
CMT	0.01	0.04	0.02	486
CLF	0.15	0.06	0.08	266
EXP	0.34	0.55	0.42	358
MNT	0	0	0	1,265
ORS	0.58	0.45	0.51	5,613
PED	0	0.02	0.01	623
ENM	0.70	0.58	0.63	696
EDP	0	0	0	15
PRV	0.04	0.05	0.04	234
RVW	0.19	0.30	0.23	289
EPA	0.30	0.40	0.34	17,130
HDG	0.46	0.20	0.28	4,358
BST	0.74	0.27	0.40	3,760
SAL	0.06	0.13	0.08	306
MNC	0.04	0.04	0.04	376
MCD	0.40	0.02	0.03	366
AAR	0.04	0.06	0.05	224
MNM	0	0	0	55
IMS	0	0	0	10
Acc.			0.26	36,678
Macro	0.20	0.20	0.17	36,678
Weighted	0.46	0.37	0.41	36,678

Table 9: Results per class for Llama-3.3-70B-Instruct

ing discourse analysis), and it is not metadiscourse about a repair. We also observe another issue in this example, which is that the label is applied to the word after the actual repair (most). This is likely due to the unidirectional nature of the model.

C Label frequency distribution

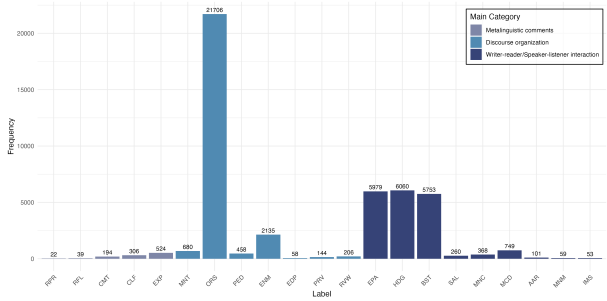

Figure 4: The distribution of MD labels in the gold-standard dataset

D Software specifications

Python: 3.11.4
numpy: 2.2.2
torch: 2.6.0+cu124
transformers: 4.48.2
All models apart from GPT-4o are available on HuggingFace:
google-bert/bert-base-cased
google-bert/bert-large-cased
SpanBERT/spanbert-base-cased
SpanBERT/spanbert-large-cased
answerdotai/ModernBERT-base
answerdotai/ModernBERT-large
FacebookAI/roberta-base
FacebookAI/roberta-large
google/electra-small-discriminator
google/electra-large-discriminator
nghuyong/ernie-2.0-base-en
nghuyong/ernie-2.0-large-en
xlnet/xlnet-base-cased
xlnet/xlnet-large-cased
meta-llama/Llama-3.3-70B-Instruct

The code can be found on this paper's associated GitHub page: https://github.com/W-Guan/Automatic-MD-annotation-with-XLNet. The dataset is available upon request, as it was collected for another study that has not been published yet. Interested parties may contact the author directly to obtain access.

E Hardware specifications

GPU: NVidia L4
GPU Memory: 24GB
CPU: AMD 9445P
Total Number of Cores: 64
Memory: 384 GB
One model tuning run of 25 epochs takes about 30 minutes (ModernBERT-base) to 1 hour (XLnet-base) on this hardware.

F Prompt template

```
You are an excellent linguist. The task is to label metadiscourse spans -
spans of words that guide the addressee through the discourse. Here are
the possible categories that you can label spans as:

RPR: Repairing - Example: "I'm sorry..."
RFL: Reformulating - Example: "to put it differently..."
CMT: Commenting - Example: "... is a difficult question"
CLF: Clarifying - Example: "I don't mean to say"
EXP: Exemplifying - Example: "for example"
ORS: Organizing statements - Example: "and", "but", "so"
PED: Providing evidentials - Example: "according to"
ENM: Enumerating - Example: "first", "at last"
EDP: Endophoric marking - Example: "As we can see in Chapter III, ..."
PRV: Previewing - Example: "We will discuss..."
RVW: Reviewing - Example: "As I said last time, ..."
EPA: Epistemic attitudes - Example: "I agree that..."
HDG: Hedges - Example: "perhaps", "might"
BST: Boosters - Example: "definitely", "should"
SAL: Speech act labels - Example: "I argue that..."
MNC: Managing comprehension - Example: "You know what I mean."
MCD: Managing channel/audience discipline - Example: "Can you hear me?"
AAR: Anticipating the audience's response - Example: "You may ask..."
MNM: Managing the message - Example: "What I want to emphasize is..."
IMS: Imagining scenarios - Example: "Suppose you are giving a speech..."

If a span should be labeled, you will annotate in the format:
@@span#LABEL##

Do not output any other text apart from the annotated input text.

Document-level examples:
{example1}
{example2}
{example3}
```

Figure 5: Prompt template for decoder LLM baseline experiment

G Hyperparameters

Model	Epochs	Batch	Alpha	Decay	Warmup steps
BERT	25	6	3.46893934104582e-05	456	0.06331306513883898
SpanBERT	25	9	1.1235689536407466e-05	583	0.03328240000998865
ModernBERT	25	4	4.27605191279545e-05	352	0.07325555301079804
RoBERTa	25	17	4.1627214448844214e-05	419	0.03135747617843722
Electra	25	4	3.8988959827328105e-05	566	0.09975738929202359
ERNIE	25	14	2.7576890378412467e-05	204	0.016803254034765108
XLNet	25	4	3.784413058653172e-05	550	0.026749712474382164

Table 10: Best hyperparameters settings for base models.