PRiSM: Enhancing Low-Resource Document-Level Relation Extraction with Relation-Aware Score Calibration

Minseok Choi **Hyesu Lim** **Jaegul Choo**
KAIST AI
{minseok.choi, hyesulim, jchoo}@kaist.ac.kr

Abstract

Document-level relation extraction (DocRE) aims to extract relations of all entity pairs in a document. A key challenge in DocRE is the cost of annotating such data which requires intensive human effort. Thus, we investigate the case of DocRE in a low-resource setting, and we find that existing models trained on low data overestimate the NA ("no relation") label, causing limited performance. In this work, we approach the problem from a calibration perspective and propose PRiSM, which learns to adapt logits based on relation semantic information. We evaluate our method on three DocRE datasets and demonstrate that integrating existing models with PRiSM improves performance by as much as 26.38 F1 score, while the calibration error drops as much as 36 times when trained with about 3% of data. The code is publicly available at https://github.com/brightjade/PRiSM.

1 Introduction

Document-level relation extraction (DocRE) is a fundamental task in natural language understanding, which aims to identify relations between entities that exist in a document. A major challenge in DocRE is the cost of annotating such documents, requiring annotators to consider relations of all possible entity combinations (Yao et al., 2019; Zaporojets et al., 2021; Tan et al., 2022b). However, there is a lack of ongoing studies investigating the low-resource setting in DocRE (Zhou et al., 2023), and we discover that most of the current DocRE models show subpar performance when trained with a small set of data. We argue that the reason is two-fold. First, the long-tailed distribution of DocRE data encourages models to be overly confident in predicting frequent relations and less sure about infrequent ones (Du et al., 2022; Tan et al., 2022a). Out of the 96 relations in DocRED (Yao et al., 2019), a widely-used DocRE dataset, the 7 most frequent relations account for 55% of the total relation triples. Under the low-resource setting, chances to observe infrequent relations become much harder. Second, DocRE models predict the NA ("no relation") label if an entity pair does not express any relation. In DocRED, about 97% of all entity pairs have the NA label. With limited data, there is a much less signal for ground-truth (GT) labels during training, resulting in models overpredicting the NA label instead.

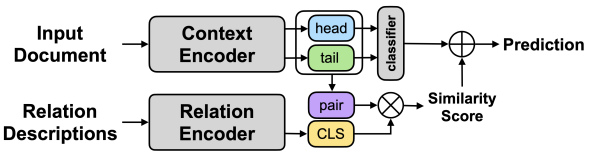

Figure 1: An overview of our proposed method. Top represents the original DocRE framework. PRiSM (bottom) leverages relation descriptions to compute scores for each relation triple. These scores are then used to reweight the prediction logits.

High confidence in common relations and the NA label and low confidence in rare relations suggest that models may be miscalibrated. We hypothesize that lowering the former and raising the latter would improve the overall RE performance. At a high level, we wish to penalize logits of frequent labels (including NA) and supplement logits of infrequent labels such that models are able to predict them without seeing them much during training. To implement such behavior, we leverage relation semantic information, which has proved to be effective in low-resource sentence-level RE (Yang et al., 2020; Dong et al., 2021; Zhang and Lu, 2022).

In this work, we propose the **P**air-**R**elation **S**imilarity **M**odule (PRiSM) that learns to adapt logits by exploiting semantic information from label descriptions, as depicted in Figure 1. Specifically, we compute a similarity function for each entity pair embedding, constructed from two entities of interest, with relation embeddings, built from corresponding label descriptions. PRiSM then learns re-

在低数据上进行的训练高估了NA标签

PRiSM（底部）利用关系描述来计算每个三重关系的分数。然后使用这些分数来重新加权预测对数。

对常见关系和NA标签的高置信度和对罕见关系的低置信度表明，模型可能被错误校准

我们假设降低前者并提高后者将提高整体关系抽取性能。

在较高层面上，我们希望对频繁标签（包括NA）的logits进行惩罚，并补充稀有标签的logits，

lation representations to output adaptive scores for each relation triple. Note that previous work mostly utilized relation representations for self-supervised learning (Dong et al., 2021; Du et al., 2022; Zhou et al., 2023), whereas PRiSM uses them to directly adjust logits, which brings a calibration effect. To elaborate further, let us say that classification logits are statistical scores and similarities are semantic scores. We have four scenarios: 1) relation is common and GT, 2) relation is common but not GT, 3) relation is uncommon but GT, and 4) relation is uncommon and not GT. In Cases 1 and 4, both statistical and semantic scores are either high or low, and thus, appending PRiSM mostly would not affect the original RE predictions. In Case 2, the statistical score is high, but the semantic score is low, possibly negative to penalize the statistical score. This is the case of PRiSM decreasing the confidence of common relations and NA label. In Case 3, the statistical score is low, but the semantic score is high, which is the case of PRiSM increasing the confidence of uncommon relations. As such, PRiSM incorporates both statistical and semantic scores such that the confidence is adjusted regardless of the relation frequency.

分类logit是统计分数

相似性是语义分数

情况1和情况4中，统计和语义分数要么都很高，要么都很低，因此，添加PRiSM主要不会影响原始的关系抽取预测。

在情况2中，统计分数很高，但语义分数很低，可能是负值以惩罚统计分数。这是PRiSM降低常见关系和NA标签置信度的情况。

在情况3中，统计分数很低，但语义分数很高，这是PRiSM增加不常见关系置信度的情况

因此，PRiSM结合了统计和语义分数，从而调整置信度，无论关系频率如何

Our technical contributions are three-fold. First, we propose PRiSM, a relation-aware calibration technique that improves model performance and adjusts model confidence on low-resource DocRE. Second, we demonstrate the performance improvement across various state-of-the-art models integrated with PRiSM. Third, we validate the effectiveness of our method on widely-used long-tailed DocRE datasets and calibration metrics.

2 Methodology

2.1 Problem Formulation

Given a document d, a set of n annotated entities $\mathcal{E} = \{e_i\}_{i=1}^{n}$, and a pre-defined set of relations $\mathcal{R} \cup \{\texttt{NA}\}$, the task of DocRE is to extract the relation triple set $\{(e_h, r, e_t) | e_h \in \mathcal{E}, r \in \mathcal{R}, e_t \in \mathcal{E}\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ from all possible relation triples, where (e_h, r, e_t) denotes that a relation r holds between head entity e_h and tail entity e_t. An entity e_i may appear k times in the document in which we denote corresponding instances as entity mentions $\{m_{ij}\}_{j=1}^{k}$. A relation r exists between an entity pair (e_h, e_t) if any pair of their mentions express the relation, and if they do not express any relation, the entity pair is then labeled as NA.

2.2 Document-Level Relation Extraction

Given a document d as an input token sequence $\boldsymbol{x} = [x_t]_{t=1}^{l}$, where l is the length of the token sequence, we explicitly locate the position of entity mentions by inserting a special token "*" before and after each mention. The presence of the entity marker has proved to be effective from previous studies (Zhang et al., 2017; Shi and Lin, 2019; Baldini Soares et al., 2019). The entity-marked document is then fed into a pre-trained language model (PLM) encoder, which outputs the contextual embeddings: $[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \text{Encoder}(\boldsymbol{x})$. We take the embedding of "*" at the start of each mention as its mention-level representation $\boldsymbol{h}_{m_{ij}}$ of the entity e_i. For extracting the entity-level representation, we apply the logsumexp pooling over all mentions $\{m_{ij}\}_{j=1}^{k}$ of the entity e_i:

$$\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{k} \exp\left(\boldsymbol{h}_{m_{ij}}\right). \tag{1}$$

The logsumexp pooling is a smooth version of max pooling and has been shown to accumulate weak signals from each different mention representation, which results in a better performance (Jia et al., 2019). We pass the embeddings of head and tail entities through a linear layer followed by non-linear activation to obtain the hidden representations: $\boldsymbol{z}_h = \tanh(\boldsymbol{W}_h \boldsymbol{h}_{e_h} + \boldsymbol{b}_h)$ and $\boldsymbol{z}_t = \tanh(\boldsymbol{W}_t \boldsymbol{h}_{e_t} + \boldsymbol{b}_t)$, where $\boldsymbol{W}_h, \boldsymbol{W}_t, \boldsymbol{b}_h, \boldsymbol{b}_t$ are learnable parameters. Then we calculate a score for relation r between entities h and t by taking a bilinear function:

$$s_{(h,r,t)} = \boldsymbol{z}_h^\top \boldsymbol{W}_r \boldsymbol{z}_t + \boldsymbol{b}_r, \tag{2}$$

where $\boldsymbol{W}_r, \boldsymbol{b}_r$ are learnable parameters.

2.3 PRiSM

Following previous work (Zhang and Lu, 2022), we feed relation descriptions to a PLM encoder to obtain the relation embedding $\boldsymbol{z}_r$ for relation r. The details of the relation descriptions used can be found in Appendix A.4. We then construct the entity pair-level representation $\boldsymbol{z}_{(h,t)}$ by mapping the head and tail embeddings to a linear layer followed by non-linear activation: $\boldsymbol{z}_{(h,t)} = \tanh(\boldsymbol{W}_{(h,t)}[\boldsymbol{z}_h; \boldsymbol{z}_t] + \boldsymbol{b}_{(h,t)})$, where $\boldsymbol{z}_h, \boldsymbol{z}_t$ are concatenated and $\boldsymbol{W}_{(h,t)}, \boldsymbol{b}_{(h,t)}$ are learnable parameters. An adaptive score for relation r between entities h and t is computed by taking a similarity function between the entity pair embedding

通过在 实体对嵌入 和 关系嵌入 之间取一个相似性函数，计算出实体h和t之间的关系r的 自适应得分
adaptive score for relation r

Model	DocRED Dev Ign F_1	DocRED Dev F_1	DocRED Test Ign F_1	DocRED Test F_1	Re-DocRED Dev Ign F_1	Re-DocRED Dev F_1	Re-DocRED Test Ign F_1	Re-DocRED Test F_1
3% training examples ($N = 100$)								
BERT$_{\text{BASE}}$	10.27 ± 1.82	10.44 ± 1.90	11.36	11.50	28.65 ± 2.87	29.40 ± 3.19	28.77 ± 3.34	29.44 ± 3.67
BERT$_{\text{BASE}}$ + **PRiSM**	**35.06** ± 0.94	**37.02** ± 0.88	**35.79**	**37.88**	**47.39** ± 0.79	**49.09** ± 0.90	**46.90** ± 1.59	**48.57** ± 1.73
RoBERTa$_{\text{BASE}}$	20.70 ± 1.91	21.31 ± 1.87	21.74	22.25	39.66 ± 2.25	40.74 ± 1.89	39.42 ± 2.80	40.53 ± 2.43
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**32.40** ± 0.85	**34.49** ± 0.76	**32.20**	**34.32**	**47.71** ± 1.03	**49.40** ± 1.14	**47.31** ± 0.96	**49.04** ± 1.05
SSAN-BERT$_{\text{BASE}}$	10.92 ± 0.88	11.18 ± 0.89	11.93	12.16	28.89 ± 1.68	29.01 ± 1.69	28.64 ± 1.89	29.29 ± 1.94
SSAN-BERT$_{\text{BASE}}$ + **PRiSM**	**32.86** ± 2.35	**34.76** ± 2.50	**34.00**	**36.03**	**46.49** ± 1.16	**48.11** ± 1.40	**46.51** ± 1.77	**48.11** ± 2.00
ATLOP-BERT$_{\text{BASE}}$	38.99 ± 2.30	40.50 ± 2.07	40.88	42.37	49.45 ± 2.09	50.60 ± 1.95	49.24 ± 2.25	50.32 ± 2.13
ATLOP-BERT$_{\text{BASE}}$ + **PRiSM**	**40.59** ± 0.68	**42.09** ± 0.66	**40.94**	**42.43**	**50.10** ± 0.53	**51.12** ± 0.64	**50.15** ± 1.11	**51.14** ± 1.17
10% training examples ($N = 305$)								
BERT$_{\text{BASE}}$	39.84 ± 0.92	41.55 ± 0.99	40.98	42.98	52.34 ± 0.66	53.54 ± 0.80	52.34 ± 0.68	53.54 ± 0.84
BERT$_{\text{BASE}}$ + **PRiSM**	**46.01** ± 0.12	**48.02** ± 0.13	**45.52**	**47.83**	**58.10** ± 0.31	**59.86** ± 0.27	**57.75** ± 0.65	**59.53** ± 0.51
RoBERTa$_{\text{BASE}}$	43.42 ± 1.09	45.20 ± 1.09	43.78	45.63	54.82 ± 1.85	56.10 ± 1.80	55.36 ± 2.18	56.67 ± 2.06
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**46.60** ± 0.20	**48.57** ± 0.29	**47.02**	**49.22**	**59.51** ± 0.36	**61.19** ± 0.32	**59.08** ± 0.61	**60.80** ± 0.52
SSAN-BERT$_{\text{BASE}}$	40.00 ± 1.62	41.65 ± 1.63	41.11	43.03	53.57 ± 0.83	54.86 ± 0.81	53.67 ± 1.55	54.94 ± 1.52
SSAN-BERT$_{\text{BASE}}$ + **PRiSM**	**46.14** ± 0.15	**48.18** ± 0.09	**45.48**	**47.72**	**58.47** ± 0.39	**60.17** ± 0.36	**58.21** ± 0.31	**59.93** ± 0.19
ATLOP-BERT$_{\text{BASE}}$	49.93 ± 1.11	51.61 ± 1.16	50.04	51.85	60.38 ± 0.46	61.52 ± 0.29	60.46 ± 0.55	61.54 ± 0.29
ATLOP-BERT$_{\text{BASE}}$ + **PRiSM**	**50.20** ± 0.68	**51.83** ± 0.64	**50.29**	**52.17**	**60.58** ± 0.18	**61.68** ± 0.17	**60.90** ± 0.37	**61.97** ± 0.40
100% training examples ($N = 3053$)								
BERT$_{\text{BASE}}$	57.15 ± 0.17	59.18 ± 0.05	57.02	59.35	71.70 ± 0.61	73.17 ± 0.55	71.01 ± 0.88	72.48 ± 0.78
BERT$_{\text{BASE}}$ + **PRiSM**	**57.82** ± 0.10	**59.93** ± 0.15	**57.17**	**59.52**	**72.92** ± 0.07	**74.25** ± 0.07	**72.35** ± 0.07	**73.69** ± 0.11
RoBERTa$_{\text{BASE}}$	58.24 ± 0.36	60.19 ± 0.38	58.00	60.10	74.00 ± 0.20	75.20 ± 0.20	73.56 ± 0.04	74.75 ± 0.04
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**58.73** ± 0.09	**60.70** ± 0.02	**58.36**	**60.51**	**74.50** ± 0.09	**75.71** ± 0.06	**74.17** ± 0.10	**75.38** ± 0.10
SSAN-BERT$_{\text{BASE}}$	57.59 ± 0.35	59.62 ± 0.24	57.71	59.79	72.59 ± 0.15	74.01 ± 0.15	71.95 ± 0.11	73.37 ± 0.11
SSAN-BERT$_{\text{BASE}}$ + **PRiSM**	**58.20** ± 0.20	**60.27** ± 0.14	**58.02**	**60.27**	**73.22** ± 0.10	**74.65** ± 0.07	**72.37** ± 0.19	**73.80** ± 0.18
ATLOP-BERT$_{\text{BASE}}$	59.22 ± 0.17	61.18 ± 0.10	**58.99**	**61.08**	72.78 ± 0.46	73.73 ± 0.37	72.60 ± 0.41	73.51 ± 0.38
ATLOP-BERT$_{\text{BASE}}$ + **PRiSM**	**59.51** ± 0.09	**61.31** ± 0.05	58.80	60.77	**72.85** ± 0.29	**73.80** ± 0.35	**72.61** ± 0.59	**73.53** ± 0.53

Table 1: Performance (%) on DocRED and Re-DocRED. Better scores between with and without PRiSM are in bold. The test results for DocRED are obtained by submitting the best dev model predictions to CodaLab[1].

Model	Macro	Macro@500	Macro@200	Macro@100
$N = 100$				
BERT$_{\text{BASE}}$	0.36 ± 0.05	0	0	–
BERT$_{\text{BASE}}$ + **PRiSM**	**7.77** ± 1.87	**4.08** ± 0.43	**0.44** ± 0.33	–
RoBERTa$_{\text{BASE}}$	1.18 ± 0.28	0	0	–
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**6.41** ± 0.77	**2.31** ± 0.82	**0.38** ± 0.28	–
$N = 305$				
BERT$_{\text{BASE}}$	9.31 ± 1.59	3.70 ± 1.46	0.29 ± 0.17	0
BERT$_{\text{BASE}}$ + **PRiSM**	**20.19** ± 0.70	**14.91** ± 0.64	**7.73** ± 0.17	**2.19** ± 1.16
RoBERTa$_{\text{BASE}}$	14.80 ± 0.51	9.13 ± 0.61	3.74 ± 0.35	0.83 ± 0.89
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**21.03** ± 0.27	**15.69** ± 0.48	**8.37** ± 0.16	**2.63** ± 1.25
$N = 3053$				
BERT$_{\text{BASE}}$	38.31 ± 0.39	34.06 ± 0.45	26.07 ± 0.72	**19.73** ± 0.96
BERT$_{\text{BASE}}$ + **PRiSM**	**38.89** ± 0.52	**34.57** ± 0.59	**26.51** ± 0.65	19.57 ± 0.71
RoBERTa$_{\text{BASE}}$	38.67 ± 1.12	34.28 ± 1.22	26.14 ± 1.44	18.69 ± 1.70
RoBERTa$_{\text{BASE}}$ + **PRiSM**	**39.12** ± 0.57	**34.72** ± 0.69	**26.45** ± 1.01	**19.23** ± 1.55

Table 2: Dev performance (%) on low-frequency relations in DocRED. Test results cannot be reported because the labels are not accessible.

and relation embedding: $s'_{(h,r,t)} = sim(\boldsymbol{z}_{(h,t)}, \boldsymbol{z}_r)$, where $sim(\cdot)$ is cosine similarity. Formally, the probability of relation r between entities h and t is simply an addition of two scores followed by sigmoid activation:

$$P(r \mid e_h, e_t) = \sigma(s_{(h,r,t)} + \lambda s'_{(h,r,t)}), \quad (3)$$

where λ is the scale factor. Finally, we optimize our model with the binary cross-entropy (BCE) loss:

$$\mathcal{L} = -\frac{1}{T} \sum_{<h,t>} \sum_{r} \mathsf{BCE}(P(r|e_h, e_t), \bar{y}_{(h,r,t)}), \quad (4)$$

where $\bar{y}$ is the target label and T is the total number of relation triples.

3 Experiments

3.1 Dataset

We evaluate our framework on three public DocRE datasets. DocRED (Yao et al., 2019) is a widely-used human-annotated DocRE dataset constructed from Wikipedia and Wikidata. Re-DocRED (Tan et al., 2022b) is a revised dataset from DocRED, addressing the incomplete annotation problem. DWIE (Zaporojets et al., 2021) is a multi-task document-level information extraction dataset consisting of news articles collected from Deutsche Welle. Dataset statistics are shown in Table 5.

3.2 Implementation Details

Our framework is built on PyTorch and Huggingface's Transformers library (Wolf et al., 2020). We use the cased BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for encoding the text and optimize their weights with AdamW (Loshchilov and Hutter, 2019). We tune our hyperparameters to maximize the F_1 score on the development set. The additional implementation details are included in Appendix B. During inference, we predict all relation triples that have probabilities higher than the F1-maximizing threshold found in the development set. We conduct our experiments with three different random seeds and report the averaged results. Following Yao et al.

[1] https://codalab.lisn.upsaclay.fr/

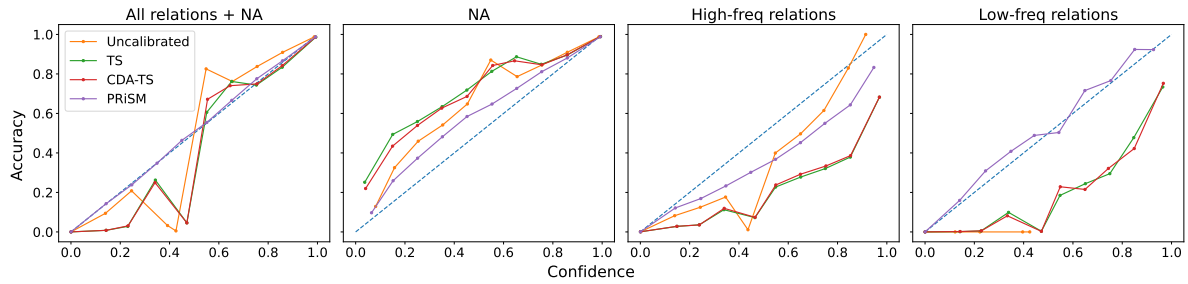

Figure 2: Reliability diagram for $\text{BERT}_{\text{BASE}}$ when trained with 3% of DocRED data.

Model	Dev F_1	Dev Macro	Test F_1	Test Macro
$N = 100$				
$\text{BERT}_{\text{BASE}}$	11.97 ± 1.78	1.79 ± 0.27	12.41 ± 2.29	1.87 ± 0.32
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	**45.20** ± 1.60	**10.34** ± 1.91	**44.31** ± 1.47	**9.47** ± 1.46
$\text{RoBERTa}_{\text{BASE}}$	50.27 ± 1.57	8.55 ± 0.56	48.29 ± 1.74	8.95 ± 0.86
$\text{RoBERTa}_{\text{BASE}}$ + **PRiSM**	**55.51** ± 1.11	**12.76** ± 2.03	**54.23** ± 1.24	**13.80** ± 0.64
$N = 305$				
$\text{BERT}_{\text{BASE}}$	52.98 ± 0.76	15.68 ± 1.71	52.05 ± 0.60	14.80 ± 0.63
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	**58.23** ± 0.40	**24.62** ± 0.59	**57.05** ± 0.23	**22.43** ± 0.88
$\text{RoBERTa}_{\text{BASE}}$	65.45 ± 1.94	21.72 ± 1.29	62.39 ± 1.29	20.39 ± 0.76
$\text{RoBERTa}_{\text{BASE}}$ + **PRiSM**	**71.18** ± 1.98	**28.36** ± 1.53	**67.12** ± 2.02	**25.82** ± 0.34
$N = 587$				
$\text{BERT}_{\text{BASE}}$	62.06 ± 0.33	25.17 ± 0.37	60.78 ± 0.25	22.93 ± 0.40
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	**66.81** ± 0.56	**28.17** ± 0.54	**66.53** ± 0.52	**29.31** ± 1.13
$\text{RoBERTa}_{\text{BASE}}$	76.23 ± 0.72	31.71 ± 0.13	74.07 ± 0.77	28.72 ± 1.54
$\text{RoBERTa}_{\text{BASE}}$ + **PRiSM**	**78.43** ± 0.12	**32.85** ± 0.37	**78.13** ± 0.61	**33.66** ± 1.24

Table 3: Performance (%) on the DWIE dataset.

(2019), all models are evaluated on F_1 and Ign F_1, where Ign F_1 excludes the relations shared by the training and development/test sets. Moreover, we measure **Macro**, which computes the average of per-class F_1, and **Macro@500**, **Macro@200**, and **Macro@100**, targeting rare relations where the frequency count in the training dataset is less than 500, 200, and 100, respectively.

3.3 Experimental Results

To simulate the low-data setting, we reduce the number of training documents N to 100 and 305, which is about 3% and 10% of the original data. To create each of the settings, we repeat random sampling until the label distribution resembles that of the full data. As shown in Table 1, we observe that performance increases consistently across different models when appended with PRiSM. Particularly, PRiSM improves performance by a large margin when trained with just 3% of data, as much as 24.43 Ign F_1 and 26.38 F_1 on the test set of DocRED for $\text{BERT}_{\text{BASE}}$. We also test PRiSM on $\text{RoBERTa}_{\text{BASE}}$ and two state-of-the-art models SSAN (Xu et al., 2021) and ATLOP (Zhou et al., 2021) and notice a similar trend, indicating that our method is effective on various existing models. We additionally evaluate PRiSM using macro metrics in Table 2 and observe that adding PRiSM improves performance on infrequent relations, especially in the low-data setting. Lastly, we validate our method on a different dataset DWIE, as illustrated in Table 3.

Method	$N = 100$ F_1(↑)	$N = 100$ ECE(↓)	$N = 100$ ACE(↓)	$N = 305$ F_1(↑)	$N = 305$ ECE(↓)	$N = 305$ ACE(↓)
Uncalibrated	10.82	0.359%	0.379%	42.56	0.137%	0.164%
TS	**38.19**	0.144%	0.173%	48.49	0.053%	0.062%
CDA-TS	37.82	0.139%	0.167%	**48.54**	0.057%	0.078%
PRiSM (ours)	37.84	**0.010%**	**0.020%**	48.10	**0.023%**	**0.020%**

Table 4: Comparison of calibration errors (with 10 bins) under a low-resource setting of DocRED.

3.4 Calibration Evaluation

We measure model calibration on two metrics: expected calibration error (ECE) (Naeini et al., 2015) and adaptive calibration error (ACE) (Nixon et al., 2019). ECE partitions predictions into a fixed number of bins and computes a weighted average of the difference between accuracy and confidence over the bins, while ACE puts the same number of predictions in each bin. We compare with general calibration methods such as temperature scaling (TS) (Guo et al., 2017) and class-distribution-aware TS (CDA-TS) (Islam et al., 2021). As reported in Table 4, PRiSM outperforms other methods in both metrics, while also maintaining a comparable RE performance. We also visualize with a reliability diagram (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005) in Figure 2. We observe that PRiSM effectively lowers the confidence of the NA label and raises the confidence of low-frequency relations (bottom 89). For high-frequency relations (top 7), confidence is adjusted in both ways. In any case, PRiSM displays the most stable, closest line to the perfect calibration (blue line).

4 Related Work

With the introduction of DocRED (Yao et al., 2019), many approaches were proposed to extract relations from a document (Wang et al., 2019; Ye et al., 2020; Zhang et al., 2021; Xu et al., 2021; Zhou et al., 2021; Xie et al., 2022). The long-tailed data problem of DocRE has been addressed in some

studies (Du et al., 2022; Tan et al., 2022a), as well as low-resource DocRE (Zhou et al., 2023); however, most require additional pretraining, which is compute- and cost-intensive, while PRiSM only requires adjusting logits in existing models. Low-resource RE has been extensively studied at the sentence level, and we specifically focus on leveraging label information (Yang et al., 2020; Dong et al., 2021; Zhang and Lu, 2022) in which PRiSM applies it to the document level. In contrast to prior work in calibration (Guo et al., 2017; Islam et al., 2021), our approach is relation-aware, updating logits at a much finer granularity.

5 Conclusion and Future Work

In this work, we propose a simple modular framework PRiSM, which exploits relation semantics to update logits. We empirically demonstrate that our method effectively improves and calibrates DocRE models where the data is long-tailed and the NA label is overestimated. For future work, we can apply PRiSM to more tasks such as event extraction and dialogue state tracking, which also enclose long-tailed data and overestimation of "null" labels.

Limitations

Although our approach is resilient to data scarcity, quite a few annotated documents are still required for the model to learn the pattern. The ultimate goal of DocRE is undoubtedly to build a model that is able to perform well on zero-shot, but we believe our approach takes a step toward that direction. Moreover, we process the long documents (> 512 tokens) in a very naive way, as described in Appendix A.3, and we think that exploration of long-sequence modeling on longer document data could further enrich the field of DocRE.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), the National Supercomputing Center with supercomputing resources including technical support (KSC-2022-CRE-0312), and Samsung Electronics Co., Ltd. We thank the anonymous reviewers for their constructive feedback.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

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.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. MapRE: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2694–2704, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yangkai Du, Tengfei Ma, Lingfei Wu, Yiming Wu, Xuhong Zhang, Bo Long, and Shouling Ji. 2022. Improving long tailed document-level relation extraction via easy relation augmentation and contrastive learning. *arXiv preprint arXiv:2205.10511*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Mobarakol Islam, Lalithkumar Seenivasan, Hongliang Ren, and Ben Glocker. 2021. Class-distribution-aware calibration for long-tailed visual recognition. *arXiv preprint arXiv:2109.05263*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale 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 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

OpenAI. 2023. ChatGPT March 23 Version.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14149–14157.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2273–2276.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14612–14620.

Wenxuan Zhou, Sheng Zhang, Tristan Naumann, Muhao Chen, and Hoifung Poon. 2023. Continual contrastive finetuning improves low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13249–13263, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Additional Dataset Details

A.1 Data Statistics

We report the statistics for the datasets in Table 5. The test set of DocRED is not included in calculating **% NA** due to its inaccessibility. 14 documents in DWIE are filtered out because of missing labels, and 1 document is removed because the annotated entities did not exist in the input document.

Statistics	DocRED	Re-DocRED	DWIE
# Train	3,053	3,053	587
# Dev	1,000	500	100
# Test	1,000	500	100
# Relation Types	97	97	66
% NA	97.05%	94.02%	97.87%

Table 5: Dataset statistics. **# Relation Types** includes the NA label. **% NA** indicates a ratio of entity pairs having the NA label over all entity pairs.

A.2 Class Distribution

We count the number of ground-truth relations in the train sets and visualize their class distribution in Figure 3. We observe that the re-annotation of Re-DocRED further skewed the class distribution, and the DWIE dataset seems to demonstrate a relatively less imbalanced distribution. Nevertheless, a few classes still exhibit high frequency, which PRiSM can handle effectively.

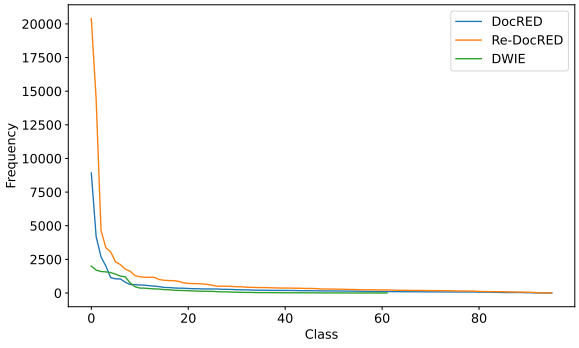

Figure 3: Dataset class distribution.

A.3 Processing Long Document

For DocRED and Re-DocRED, most of the documents contain less than 512 tokens, and thus we follow the previous work and truncate all of the inputs to 512 tokens, which is the maximum sequence length of BERT. However, we notice that the DWIE dataset mostly contains documents much longer than 512 tokens (as shown in Figure 4) in which the truncation hurts the performance significantly. Therefore, we choose the most naive way of splitting the input document into multiple chunks of length 512 and passing them through the encoder multiple times. The performance improvement over the truncation method is demonstrated in Table 6.

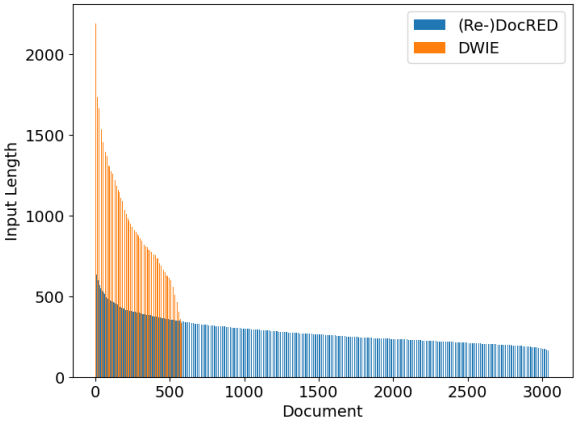

Figure 4: Lengths of input documents in training sets.

Method	Dev F_1	Test F_1
$N = 100$		
Truncation	36.53 ± 2.09	35.39 ± 0.71
Chunking (ours)	$\mathbf{45.20} \pm 1.60$	$\mathbf{44.31} \pm 1.47$
$N = 305$		
Truncation	53.65 ± 0.63	51.30 ± 0.62
Chunking (ours)	$\mathbf{58.23} \pm 0.40$	$\mathbf{57.05} \pm 0.23$
$N = 587$		
Truncation	61.34 ± 0.52	59.17 ± 1.86
Chunking (ours)	$\mathbf{66.81} \pm 0.56$	$\mathbf{66.53} \pm 0.52$

Table 6: Performance comparison of long document processing methods. The model is fixed with $\text{BERT}_{\text{BASE}}$ + PRiSM evaluating on the DWIE dataset.

A.4 Relation Descriptions

We provide a small set of relations and their descriptions in DocRED and DWIE in Table 7 and 8. For DocRED, a full list can be found either in their paper (Yao et al., 2019) or link[2]. For DWIE,

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

a full list is not available publicly; however, we were able to obtain a draft of the annotation documentation from the author. Unannotated relation descriptions were crafted with the help of a large language model (OpenAI, 2023).

Relation Name	Description
head of government	head of the executive power of this town, city, municipality, state, country, or other governmental body
country	sovereign state of this item; don't use on humans
place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
country of citizenship	the object is a country that recognizes the subject as its citizen
member of sports team	sports teams or clubs that the subject currently represents or formerly represented

Table 7: DocRED relation descriptions.

Relation Name	Description
based_in0	Relations between organizations and the countries they are based in
in0	Relations between geographic locations and the countries they are located in
citizen_of	Relations between people and the country they are citizens of
based_in0-x	Relations between organizations and the nominal variations of the countries they are based in
citizen_of-x	Relations between people and the nominal variations of the countries they are citizens of

Table 8: DWIE relation descriptions.

B Additional Details for PRiSM

B.1 Detailed Experimental Setup

Device. For all our experiments, we trained the networks on a single NVIDIA TITAN RTX GPU with 24GB of memory.

Model Size. PRiSM shares parameters with the PLM used when learning relation representations. The only additional parameter weights come from a linear layer constructing pair representations and an extra embedding space initialized for relation tokens. The number of trainable parameters for each model is illustrated in Table 9.

Model	# Parameters
$\text{BERT}_{\text{BASE}}$	108,310,272
$\text{BERT}_{\text{BASE}}$-DocRE	114,259,297
$\text{BERT}_{\text{BASE}}$-DocRE + **PRiSM**	115,514,209
$\text{RoBERTa}_{\text{BASE}}$	124,645,632
$\text{RoBERTa}_{\text{BASE}}$-DocRE	130,594,657
$\text{RoBERTa}_{\text{BASE}}$-DocRE + **PRiSM**	131,849,569

Table 9: Comparison of model parameters. **DocRE** includes a bilinear layer and two linear layers for constructing head and tail representations.

GPU Hours. Adding PRiSM takes a slightly longer computation time than the existing DocRE models due to having to pass the PLM twice. Note that PRiSM is built for a low-resource setting in which the computation time does not seem to differ as much. The comparison of GPU hours is reported in Table 10.

Model	Training Hours
3% training examples	
$\text{BERT}_{\text{BASE}}$	0.8
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	0.8
10% training examples	
$\text{BERT}_{\text{BASE}}$	0.8
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	1.0
100% training examples	
$\text{BERT}_{\text{BASE}}$	2.7
$\text{BERT}_{\text{BASE}}$ + **PRiSM**	3.5

Table 10: GPU hours for DocRED training. Time for evaluating and saving the model every epoch is included.

Hyperparameters. We perform a grid search on finding the best hyperparameter configuration and report the tuning range used for our experiments in Table 11. The evaluation on the validation set is performed for every epoch and the tolerance increases by 1 when the validation F_1 is worse than the previous evaluation. The training stops early when the count reaches the max tolerance.

Model	Dataset	Hyperparameter	Range	Best
$\text{BERT}_{\text{BASE}}$, $\text{RoBERTa}_{\text{BASE}}$	(Re-)DocRED	batch size	{ 4, 8 }	4
		learning rate	{ 1e-5, 2e-5, 3e-5, 4e-5, 5e-5, 1e-4 }	3e-5
		warmup ratio	{ 0, 0.06, 0.1 }	0.06
		λ	{ 1, 5, 10, 100 }	10
		max grad norm	{ 1.0 }	1.0
		max tolerance	{ 5 }	5
		epoch	{ 30 }	30
$\text{BERT}_{\text{BASE}}$, $\text{RoBERTa}_{\text{BASE}}$	DWIE	batch size	{ 4, 8 }	4
		learning rate	{ 1e-5, 2e-5, 3e-5, 4e-5, 5e-5, 1e-4 }	5e-5
		warmup ratio	{ 0, 0.06, 0.1 }	0.06
		λ	{ 1, 5, 10, 100 }	10
		max grad norm	{ 1.0 }	1.0
		max tolerance	{ 5 }	5
		epoch	{ 30 }	30

Table 11: Hyperparameter tuning range and best values used in the experiments.

B.2 Evaluation Details

We elaborate on the details of calculating calibration errors. We utilize two metrics in our paper. ECE (Naeini et al., 2015) divides the probability interval into a fixed number of bins, calculates the difference between the accuracy of the predictions and the mean of the probabilities (confidence) in each bin, and computes a weighted average over

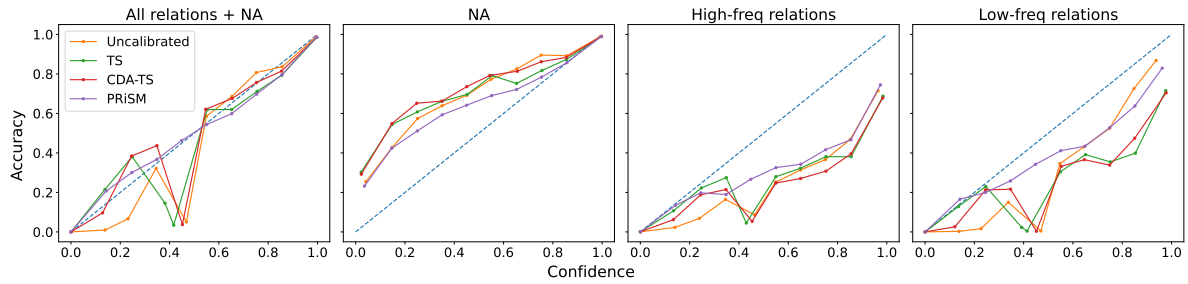

Figure 5: Reliability diagram for $\text{BERT}_{\text{BASE}}$ when trained with 10% of DocRED data.

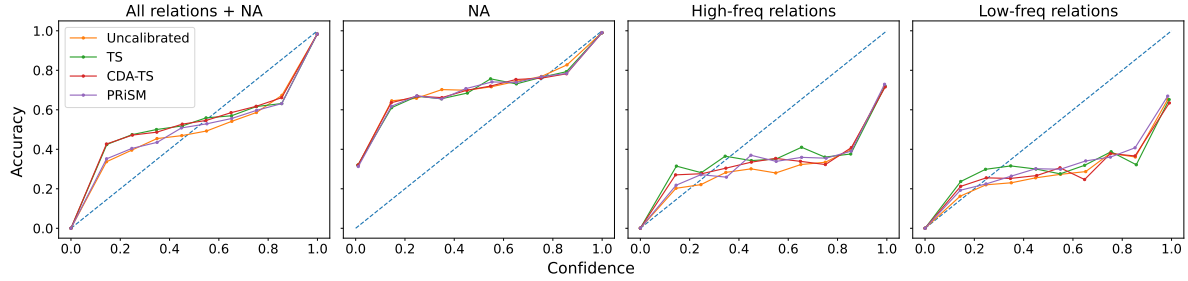

Figure 6: Reliability diagram for $\text{BERT}_{\text{BASE}}$ when trained with 100% of DocRED data.

the bins. Formally, the equation can be written as

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{T} |\text{acc}(b) - \text{conf}(b)|, \qquad (5)$$

where n_b is the number of predictions in bin b, B is a hyperparameter for the total number of bins, and T is the total number of samples. On the other hand, ACE (Nixon et al., 2019) divides up the probability interval by having the same number of predictions in each bin, thereby mitigating the issue of only calibrating the most confident samples. The equation is written as

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |\text{acc}(r,k) - \text{conf}(r,k)|, \quad (6)$$

where $\text{acc}(r,k)$ and $\text{conf}(r,k)$ are the accuracy and confidence of adaptive calibration range r for class label k, respectively.

B.3 Additional Calibration Results

We visualize the calibration of the rest of the data setting (i.e., 10% and 100% training data) with reliability diagrams in Figure 5 and 6. We notice that PRiSM is still effective with 10% of training data, but with full data, the performance gain is minimal; that is, the line barely moves toward the perfect calibration line.

We understand that calibration results on models other than the $\text{BERT}_{\text{BASE}}$ may be important in demonstrating the effectiveness of PRiSM. As shown in Table 12, we find that $\text{RoBERTa}_{\text{BASE}}$ and $\text{SSAN-BERT}_{\text{BASE}}$ follow the same trend as $\text{BERT}_{\text{BASE}}$, showing the lowest calibration error when PRiSM is appended. We also observe a similar pattern with the Re-DocRED data. We do not report results for ATLOP because the calibration errors for ATLOP must be computed differently, as it does not use probabilities (confidence) for prediction.

	$N = 100$			$N = 305$		
Method	**F_1(↑)**	**ECE(↓)**	**ACE(↓)**	**F_1(↑)**	**ECE(↓)**	**ACE(↓)**
DocRED results						
BERT	10.82	0.359%	0.379%	42.56	0.137%	0.164%
BERT + **PRiSM**	**37.84**	**0.010%**	**0.020%**	**48.10**	**0.023%**	**0.020%**
RoBERTa	22.55	0.671%	0.691%	45.83	0.237%	0.259%
RoBERTa + **PRiSM**	**35.10**	**0.015%**	**0.025%**	**48.70**	**0.022%**	**0.020%**
SSAN-BERT	11.93	0.368%	0.390%	42.82	0.128%	0.152%
SSAN-BERT + **PRiSM**	**36.96**	**0.019%**	**0.019%**	**48.28**	**0.023%**	**0.023%**
Re-DocRED results						
BERT	32.75	0.367%	0.407%	54.44	0.185%	0.190%
BERT + **PRiSM**	**49.90**	**0.038%**	**0.048%**	**60.17**	**0.056%**	**0.036%**
RoBERTa	42.40	0.722%	0.718%	57.61	0.191%	0.208%
RoBERTa + **PRiSM**	**50.34**	**0.055%**	**0.051%**	**61.55**	**0.047%**	**0.033%**
SSAN-BERT	30.28	0.362%	0.407%	55.60	0.125%	0.148%
SSAN-BERT + **PRiSM**	**49.04**	**0.050%**	**0.053%**	**60.57**	**0.051%**	**0.036%**

Table 12: Comparison of calibration errors (with 10 bins) of different models under a low-resource setting of DocRED and Re-DocRED.