Towards Robust Few-Shot Relation Classification: Incorporating Relation Description with Agreement

Mengting Hu[1] **Hang Gao**[2] * **Jianfeng Wu**[3] **Ming Jiang**[3] **Yalan Xie**[3]
Zhunheng Wang[3] **Rui Ying**[3] **Xiaoyi Liu**[3] **Ruixuan Xu**[3] **Renhong Cheng**[3]
[1] College of Software, Nankai University
[2] College of Artificial Intelligence, Tianjin University of Science and Technology
[3] College of Computer Science, Nankai University
mthu@nankai.edu.cn, wjf@mail.nankai.edu.cn

Abstract

Few-shot relation classification aims to recognize the relation between two mentioned entities, with the help of only a few support samples. However, a few samples tend to be limited for tackling unlimited queries. If a query cannot find references from the support samples, it is defined as none-of-the-above (NOTA). Previous works mainly focus on how to distinguish $N + 1$ categories, including N known relations and one NOTA class, to accurately recognize relations. However, the robustness towards various NOTA rates, i.e. *the proportion of NOTA among queries*, is under investigation. In this paper, we target the robustness and propose a simple but effective framework. Specifically, we introduce relation descriptions as external knowledge to enhance the model's comprehension of the relation semantics. Moreover, we further promote robustness by proposing a novel agreement loss. It is designed for seeking decision consistency between the instance-level decision, i.e. *support samples*, and relation-level decision, i.e. *relation descriptions*. Extensive experimental results demonstrate that the proposed framework outperforms strong baselines while being robust against various NOTA rates. The code is released on GitHub at https://github.com/Pisces-29/RoFRC.

1 Introduction

Few-shot relation classification (FSRC) is a popular task in the information extraction field. It aims to predict the relation between two entities in a sentence, by only referencing a few known samples. Previous FSRC models are usually trained on a collection of meta-tasks sampled from the training corpus (Liu et al., 2022c). A meta-task can also be called an *"episode"*, which contains a support set and a query set. The support set comprises N non-overlapping categories, and each category consists of K instances. Each query is classified with the help of the support samples. By learning meta-knowledge from these episodic training tasks, the model can quickly adapt to classifying new relations with only a few examples.

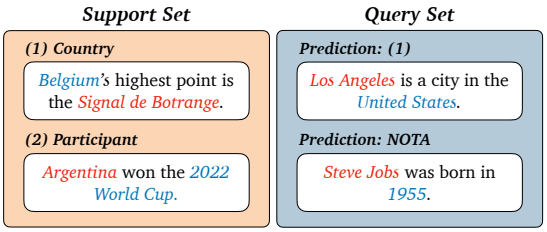

Figure 1: An example for a 2-way 1-shot episode with 50% NOTA rate. The support set involves 2 classes, each containing only 1 instance. The head entity and tail entity are marked in red and blue, respectively. The query set contains two instances that need to be predicted.

As an example shown in Figure 1, the support set provides two known relations, i.e. *"country"* and *"participant"*. The first query instance, which expresses a *"country"* relation, can be easily classified by referring to the support set. But this is an ideal situation. However, in real-world applications, the FSRC model often encounters query instances with relations that are not present in the support set. For the second query, its relation between *"Steve Jobs"* and *"1955"* is *"born in"*. It is unable to find a reference from the support set. Gao et al. (2019) define such relation as *none-of-the-above* (NOTA) and first propose that the FSRC model should involve $N + 1$ categories, including N known relations and an additional NOTA category. By introducing the NOTA relation, the model can more robustly handle the open-ended nature of real-world relation extraction tasks, where novel relations are likely to be encountered.

To address this problem, previous works mainly aim to distinguish $N + 1$ classes by introducing learnable vectors (Sabo et al., 2021), formulating a multiple-choice problem (Liu et al., 2022a), or gen-

* Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8335–8349
November 4-9, 2025

erating NOTA representations (Liu et al., 2022b). While these approaches have shown promising results, their robustness towards different NOTA rates remains an open question. Most existing works evaluate their models using fixed NOTA rates, such as 30% or 50% (Liu et al., 2022a). This means that the proportion of query instances with the NOTA relation is kept constant during training and testing. However, this assumption may not hold in practical applications, where the NOTA rate can vary significantly across different groups of queries. The sensitivity of FSRC models to NOTA rates has important implications for their real-world applicability.

To deal with this issue, we propose a simple but effective architecture, named Robust Few-shot Relation Classification (**RoFRC**), which improves not only the model's robustness in varying NOTA rate scenarios but also its overall performance. Initially, except for the instance-level decision (a.k.a. using support samples), we introduce relation descriptions as external knowledge for the relation-level decision to assist the model in accurately understanding the semantics of relations. For example, the description of the "*country*" relation is "*sovereign state of this item*". Assuming we train a model with the 50% NOTA rate scenario in Figure 1, it faces an unseen relation "*province*". By leveraging the description "*a territory governed as a unit of a country or empire*", our method gains an in-depth comprehension of semantics and reduces the negative effects of the increase or decrease of the NOTA proportion in the query set.

Additionally, differing from previous work that incorporates relation description for encoding representations (Liu et al., 2022b), we leverage it into the decision stage. To further promote robustness, we propose a novel agreement loss to seek decision consistency between the instance-level and relation-level decisions. Such consistency forces the proposed RoFRC to comprehend relation semantics better. Our method is evaluated on the popular FewRel (Han et al., 2018) and NYT-25 dataset, and extensive experiments show its effectiveness and robustness in handling varying NOTA rate scenarios, outperforming state-of-the-art few-shot relation classification approaches.

In summary, the main contributions of our work are as follows:

- Our proposed RoFRC architecture introduces a relation-level decision to facilitate the model's robustness towards various NOTA rates.
- Additionally, a novel agreement loss function is introduced during training to ensure the consistency of instance-level and relation-level decisions.
- Furthermore, the proposed method is evaluated via vastly compared with strong baseline models, along with the popular large language models. Extensive experimental results show its superiority and robustness.

2 Related Work

Previous research (Kumar, 2017; Zhang et al., 2018) in relation classification typically train models on labeled datasets with a predetermined number of classes, limiting their capability to handle unseen relations. To address this issue, Han et al. (2018) propose the use of few-shot relation classification tasks and introduce a comprehensive supervised dataset called FewRel. Liu et al. (2022c) propose a simple and efficient framework that incorporates relation description information, proving to be a significant improvement over HCRP (Han et al., 2021). Borchert et al. (2024) present a novel method that jointly leverages contrastive learning and diverse sentence representations. Dong et al. (2024) propose a method that incorporates relation-aware prompt templates and multi-level contrastive learning to improve prototype representations and mitigate relation confusion. Sun and Chen (2025) propose a Local-to-Global Optimization framework that enhances prototype learning through entity-relation alignment, local contrastive learning, and a local adaptive focal loss. In terms of incorporating external information, Wang et al. (2020) propose the incorporation of additional relative position and syntactic relation information. Yang et al. (2020) utilize a collaborative attention mechanism to integrate text descriptions of relations and entities. Furthermore, Yang et al. (2021) introduce the utilization of inherent entity concepts to provide clues for relation classification.

Gao et al. (2019) identify the NOTA challenge in the few-shot relation classification task, which simulates real-world conditions. They propose a solution to the NOTA problem by utilizing a sentence-pair model. Sabo et al. (2021) represent the prototype of the NOTA relation in a prototypical network by using multiple trainable vectors. Liu et al.

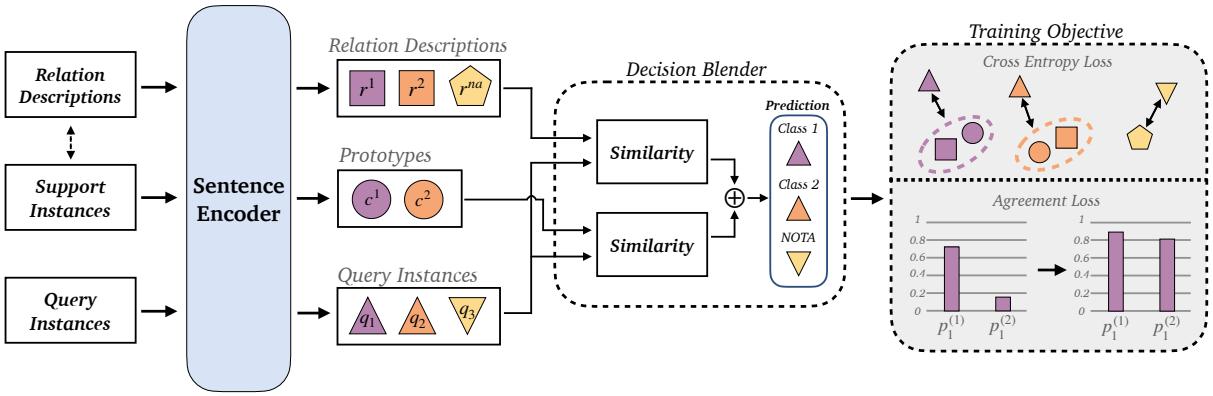

Figure 2: The overview of the RoFRC framework. $\bigoplus$ refers to the addition of the two similarities in a corresponding relation. In this example, the support set is 2-way. Both relation descriptions and instances are encoded using the same encoder. In addition, there is a special NOTA relation description in the relation set.

(2022a) propose a multiple-choice network with the pretraining fine-tuning paradigm for the few-shot relation classification task, in which an $N+1$ option for handling NOTA challenge. Wu et al. (2023) design a NOTA detection model based on instance density for classifying NOTA instances after the prototypical network. Additionally, Liu et al. (2022b) are the first to introduce relation descriptions to address the NOTA challenge. Their novel rectification module fuses relation descriptions with prototypes to generate rectified prototypes and a pseudo-NOTA prototype for $(N+1)$-way classification. Compared to PRM, our proposed method aims to enable the model to learn how to refer to support instances and fully comprehend the actual meaning of relations. Therefore, RoFRC exhibits robustness in addressing the NOTA challenge.

3 Methodology

3.1 Task Definition

The training set $\mathcal{D}_{train}$ is a collection of samples for training, denoted as (x, e_h, e_t, r), where x represents a natural language sentence, e_h and e_t represent the head and tail entity of the sentence respectively, and r denotes the relation between the pair of entities. To accomplish few-shot relation classification, we utilize the N-way K-shot setting to construct multiple episodes from the training set $\mathcal{D}_{train}$. This approach is commonly referred to as the meta-learning paradigm (Vilalta and Drissi, 2002; Vanschoren, 2018). Each episode contains the following components: $\{\mathcal{S}, \mathcal{Q}, \mathcal{R}\}$. The relation set $\mathcal{R} = \{r^1, r^2, ..., r^N\}$ is a collection of N non-overlapping relations, which are randomly sampled from $\mathcal{D}_{train}$. For each relation r^i, the support set $\mathcal{S}$ contains K instances of that relation category, randomly sampled from the training set $\mathcal{D}_{train}$.

$$\mathcal{S} = \{(s_j^i, r_j^i)\} \quad i \in [1, N], j \in [1, K] \tag{1}$$

where s_j^i denotes the j-th instance belonging to the relation r^i.

The query set $\mathcal{Q}$ consists of M query instances. Every query instance belongs to either one of the N known categories or NOTA.

$$\mathcal{Q} = \{(q_j, r_j)\} \quad j \in [1, M] \tag{2}$$

where q_j is the j-th query instance. If the relation of q_j is not included in set $\mathcal{R}$, it will be classified as the NOTA category r^{na}. The model is optimized by sampling episodes from the $\mathcal{D}_{train}$, allowing it can quickly adapt to new tasks arising from the unseen relations of the testing set $\mathcal{D}_{test}$.

The overall architecture of RoFRC is shown in Figure 2. Firstly, the sentence encoder aims to transform relations and instances into vectors. Then the encoded representations are fed into the decision blender to aggregate relation-level and instance-level decisions. Lastly, the training objective optimizes model parameters and ensures consistency between relation-level and instance-level decisions.

3.2 Sentence Encoder

We use BERT (Devlin et al., 2019) as the encoder to obtain embeddings, which is shared by three inputs, including *relation descriptions*, *support instances* to obtain *prototypes*, and *query instances*.

Relation Descriptions: For each relation, we construct a context sequence by concatenating its name and description as [[CLS], *relation name,* [SEP], *relation description*]. For example, [[CLS], *country,* [SEP], *sovereign state of this item*]. We introduce the "*none-of-the-above*" relation name for NOTA. Additionally, its relation description is defined as "*the relation of the query is not mentioned in the relation set*". This context sequence is then fed into the BERT encoder. We follow the approach proposed by Liu et al. (2022c) to obtain the representation of each relation r^i. Specifically,

$$\boldsymbol{r^i} = \mathrm{E}(r^i) = [\boldsymbol{r^i_{view1}}; \boldsymbol{r^i_{view2}}] \tag{3}$$

where $\boldsymbol{r^i_{view1}} \in \mathbb{R}^d$ is the embedding of the [CLS] token, $\boldsymbol{r^i_{view2}} \in \mathbb{R}^d$ is the average embedding of all tokens, $[\cdot\,;\,\cdot]$ represents the concatenation operation and $\mathrm{E}(\cdot)$ indicates the BERT encoder.

Prototypes: Each support instance is tokenized as a context token sequence $s^i_j = [x_0, x_1, ...x_n]$, where $x_0 =$ [CLS] is a special token that represents the beginning of the sequence. The head entity e_h in the sequence is surrounded by two special tokens, [unused0] and [unused2], while the tail entity e_t is enclosed by [unused1] and [unused3]. Following Baldini Soares et al. (2019), the representation of a support instance is obtained by concatenating the hidden states corresponding to start tokens [unused0] and [unused1] of two entities.

$$\boldsymbol{s^i_j} = \mathrm{E}(s^i_j) \tag{4}$$

where $\boldsymbol{s^i_j} \in \mathbb{R}^{2d}$. We then obtain the prototype $\boldsymbol{c^i}$ of each relation r^i by averaging the embeddings of K instances in the support set. Specifically,

$$\boldsymbol{c^i} = \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{s^i_j} \quad i \in [1, N] \tag{5}$$

where $\boldsymbol{c^i} \in \mathbb{R}^{2d}$. The prototype $\boldsymbol{c^i}$ can be regarded as a semantic summary of K instances from the relation r^i.

Query Instances: A query instance q_j is also encoded as $\boldsymbol{q_j} = \mathrm{E}(q_j)$ using Eq. (4). We employ the identical BERT encoder to encode both instances and relations, facilitating their embedding in a shared semantic space.

3.3 Decision Blender

Our approach introduces relation-level decision and blends it with the instance-level decision, which enables the model to efficiently handle the few-shot relation classification task and detect NOTA accurately.

Instance-Level Decision: We first compute the similarity α^i_j between the query instance $\boldsymbol{q_j}$ and each relation prototype $\boldsymbol{c^i}$.

$$\alpha^i_j = \boldsymbol{q_j} \odot \boldsymbol{c^i} \tag{6}$$

where $\odot$ represents the vector dot operation. Usually, a larger similarity score indicates more possibility that a query sample mentions that relation. This way can be regarded as the instance-level decision since the semantics of each relation are represented with support instances. However, there is no support instance corresponding to the NOTA category. In order to tackle this problem, we introduce the relation-level decision as an auxiliary.

Relation-Level Decision: To make the relation-level decision, we measure the similarity β^i_j between the query instance $\boldsymbol{q_j}$ and each relation description vector $\boldsymbol{r^i}$.

$$\beta^i_j = \boldsymbol{q_j} \odot \boldsymbol{r^i} \tag{7}$$

Comprehending the relation semantics contributes to better distinguishing them. Except for the description representations of known relations $\boldsymbol{r^i}, i \in [1, N]$, the description vector for the NOTA category is denoted as $\boldsymbol{r^{na}}$. Therefore, the similarity β^{na}_j between a query instance and the NOTA relation can be defined as follows:

$$\beta^{na}_j = \boldsymbol{q_j} \odot \boldsymbol{r^{na}} \tag{8}$$

Blending Decisions: Intuitively, two levels of decisions are both important to provide relation semantics. We mix them up by simply calculating the average of their similarities with both the prototype and relation description.

$$\gamma^i_j = \frac{\alpha^i_j + \beta^i_j}{2} \tag{9}$$

Deriving the final decision still meets an issue since there is no corresponding prototype for the NOTA relation. The similarity between a query instance and the NOTA relation only depends on the relation-level decision. Thus, we design the final predicted probability by blending decisions as follows. Assuming that the similarities used for prediction concatenated as $\boldsymbol{\delta_j} = \{\gamma^1_j, ..., \gamma^N_j, \beta^{na}_j\}$.

$$\mathcal{P}(r_j|q_j) = \mathrm{softmax}(\boldsymbol{\delta_j}) \tag{10}$$

Overall, RoFRC conducts a $(N+1)$-category prediction by selecting the most similar relation. N known relations and NOTA are treated separately. Concretely, the known N relations are constrained by the two levels of similarity. If a query instance still shows higher similarity to NOTA than the N known relations, it is obviously appropriate to predict the query instance as not mentioned above.

3.4 Training Objectives

Cross-Entropy Loss: Our method applies the cross-entropy loss function to lessen the negative log probability of the query instance and its corresponding ground-truth category. The cross-entropy loss $\mathcal{L}_c$ is as follows:

$$\mathcal{L}_c = -\sum_{j=1}^{M} \log \mathcal{P}(r_j|q_j) \tag{11}$$

where M is the batch size, and $\mathcal{P}(r_j|q_j)$ is the probability that the model predicts that the query instance q_j belongs to the true category r_j, as shown in equation (10).

Agreement Loss: Although equal importance is assigned to the prototype similarity α_j^i and relation description similarity β_j^i when computing predicted probabilities for the known N classes, there is no constraint between them. This might result in one decision dominating the other during practice. To address this problem, we propose a novel agreement loss $\mathcal{L}_{agree}$, which is inspired by Pagliardini et al. (2022). This loss aims to ensure the consistency between two decisions.

According to Vazhentsev et al. (2022), training neural networks with softmax layer and cross-entropy loss would make the output distribution overconfident and peaked. Therefore, in the agreement loss $\mathcal{L}_{agree}$, we pay special attention to the largest similarity. Concretely, assume a query instance q_j with N prototype similarities represented as are $\boldsymbol{\alpha_j} = \{\alpha_j^1, \alpha_j^2, ..., \alpha_j^N\}$. Its maximum probability is defined as $p_j^{(1)}$.

$$p_j^{(1)} = \max(\text{softmax}(\boldsymbol{\alpha_j})) \tag{12}$$

$$m = \text{argmax}(\text{softmax}(\boldsymbol{\alpha_j})) \tag{13}$$

where m indicates the relation with the maximum probability, denoted as r^m.

Then, m is exploited to choose the corresponding probability from the description similarities $\boldsymbol{\beta_j} = \{\beta_j^1, \beta_j^2, ..., \beta_j^N\}$.

$$p_j^{(2)} = \text{softmax}(\boldsymbol{\beta}_j)[m] \tag{14}$$

With the selected probabilities from two levels of decisions, the proposed agreement loss $\mathcal{L}_{agree}$ is formulated as:

$$\mathcal{L}_{agree} = -\sum_{j=1}^{M} \log(p_j^{(1)} \cdot \overline{p}_j^{(2)} + \overline{p}_j^{(1)} \cdot p_j^{(2)}) \tag{15}$$

where $\overline{p}_j^{(1)} = 1 - p_j^{(1)}$ and $\overline{p}_j^{(2)} = 1 - p_j^{(2)}$. While attempting to maximize $\mathcal{L}_{agree}$, the prototype and relation similarity of non-NOTA categories gradually become closer in order to ensure decision consistency. Here, the probability $p^{(1)}$ (or $\overline{p}^{(1)}$) and $\overline{p}^{(2)}$ (or $p^{(2)}$) pull each other and finally reach a trade-off balance. Thus, this loss follows an adversarial manner.

Joint Training Objective: Ultimately, our model is optimized by jointly combining two training objectives, which minimizes the cross-entropy loss and maximizes the agreement loss.

$$\mathcal{L} = \mathcal{L}_c - \lambda \mathcal{L}_{agree} \tag{16}$$

where λ is a hyper-parameter to balance the contribution of $\mathcal{L}_{agree}$.

4 Experiments

4.1 Baseline Methods

The following baseline methods are chosen for comparison. **O-Proto** (Tan et al., 2019) is based on prototypical networks (Snell et al., 2017) and uses cosine similarity to solve the few-shot out-of-domain detection challenge. **BERT-Pair** (Gao et al., 2019) uses a sentence pairing model to concatenate a query instance and support instances to generate a similarity score to indicate whether they share the same relation. **MNAV** (Sabo et al., 2021) also employs the prototypical network, where the learnable vectors represent the NOTA class, and a query instance is classified based on its similarity with the NOTA vectors in the embedding space. **MCMN** (Liu et al., 2022a) converts all candidate relation names into multiple-choice prompts and it adds an extra option for detecting NOTA. **PRM** (Liu et al., 2022b) designs a rectification module and uses relation description to generate a rectified NOTA prototype for $N+1$ classification. **DProto** (Wu et al., 2023) analyzes the density difference

Methods	15%NOTA		30%NOTA		50%NOTA		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
5-way 1-shot								
O-Proto	71.91±1.98	70.62±2.02	72.48±1.21	71.30±1.70	73.69±0.72	69.87±0.85	72.69	70.60
BERT-Pair	75.81±1.30	74.70±1.33	76.87±1.05	75.81±1.25	78.94±0.81	75.12±1.01	77.21	75.21
MNAV	77.03±1.91	75.92±1.96	76.86±1.34	76.33±1.58	76.55±2.00	74.21±1.55	76.81	75.49
MCMN	83.75±2.01	79.61±2.51	83.27±1.53	80.33±2.16	82.36±2.24	79.06±1.82	83.13	79.67
PRM	80.38±1.61	79.54±1.67	82.10±1.24	81.15±1.61	84.99±0.41	81.30±1.00	82.49	80.80
RoFRC	**84.64±1.69**	**83.92±1.31**	**85.59±0.91**	**85.05±1.07**	**87.07±0.65**	**84.69±0.65**	**85.77**	**84.55**
5-way 5-shot								
O-Proto	81.38±1.26	80.45±1.29	81.42±1.11	81.04±1.19	81.38±0.96	79.39±1.06	81.38	80.29
BERT-Pair	82.83±0.62	81.98±0.62	83.76±0.46	83.00±0.54	85.49±0.29	82.62±0.45	84.02	82.53
MNAV	85.05±0.97	84.22±1.03	84.36±1.34	84.45±1.15	82.91±2.61	82.15±1.85	84.10	83.60
MCMN	**87.56±1.93**	84.04±2.52	83.16±3.37	83.40±3.08	75.07±5.79	79.14±3.93	81.93	82.19
PRM	84.86±1.98	84.17±2.01	86.01±1.50	85.40±1.80	88.00±0.55	85.46±1.16	86.29	85.01
DProto	85.37±0.61	84.67±0.61	84.87±0.71	84.90±0.66	83.05±1.03	82.66±1.10	84.43	84.08
RoFRC	87.38±0.47	**86.75±0.47**	**87.76±0.35**	**87.44±0.38**	**88.78±0.44**	**87.06±0.32**	**87.97**	**87.08**

Table 1: Evaluation results of baseline methods and the proposed RoFRC, in terms of accuracy (%) and F1 (%), on the FewRel dataset. The reported results are the average and standard deviation of five runs.

Methods	15%NOTA		30%NOTA		50%NOTA		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
5-way 1-shot								
O-Proto	39.71±2.83	36.64±2.94	40.58±2.32	37.14±2.66	42.11±1.75	35.39±2.23	40.80	36.39
BERT-Pair	47.35±2.00	44.53±2.07	48.30±1.33	44.85±1.73	50.08±1.90	43.07±1.35	48.58	44.15
MNAV	40.46±3.47	37.48±3.45	40.68±3.47	37.56±3.45	41.03±4.62	35.39±3.35	40.72	36.81
MCMN	**66.17±3.01**	59.05±3.39	65.28±1.87	59.41±3.09	63.54±2.03	57.35±2.37	65.00	58.59
PRM	61.60±3.87	59.51±3.92	61.26±3.00	59.46±3.56	60.88±2.83	56.81±3.01	61.25	58.60
RoFRC	65.55±1.83	**63.66±1.93**	**65.31±1.64**	**63.99±1.78**	**65.11±1.55**	**61.31±1.71**	**65.32**	**62.99**
5-way 5-shot								
O-Proto	52.55±1.54	49.74±1.55	50.49±1.50	48.94±1.45	46.74±2.00	44.98±1.60	49.93	47.89
BERT-Pair	58.76±3.06	56.45±3.26	59.32±2.83	56.88±3.23	60.28±3.30	54.78±3.11	59.45	56.03
MNAV	54.85±2.37	51.67±2.44	51.08±2.34	50.10±2.45	44.27±2.34	44.96±2.37	50.07	48.91
MCMN	**71.96±0.39**	65.34±0.46	68.26±0.44	64.71±0.23	61.64±2.06	60.94±0.49	67.29	63.66
PRM	65.00±4.88	63.28±5.05	65.26±4.05	63.60±4.80	65.85±4.61	61.59±4.30	65.37	62.82
DProto	57.36±2.60	54.70±2.72	53.93±2.47	53.22±2.59	45.68±2.19	47.04±2.26	52.32	51.65
RoFRC	70.04±1.55	**68.45±1.51**	**69.35±1.07**	**68.55±1.27**	**67.93±2.42**	**65.76±1.53**	**69.10**	**67.59**

Table 2: Evaluation results of baseline methods and the proposed RoFRC, in terms of accuracy (%) and F1 (%), on the NYT-25 dataset. The reported results are the average and standard deviation of five runs.

between non-NOTA instances and NOTA instances, and adds a density detection module after the prototypical network to classify NOTA instances.

4.2 Overall Results

The experimental results are shown in Table 1 and Table 2. Under identical experimental settings, our proposed method outperforms other strong baselines significantly. We first focus on the experimental results of the FewRel dataset in Table 1. In the 5-way 1-shot setting, the average accuracy of RoFRC improves by 2.64% compared to MCMN. Furthermore, it achieves an improvement of 4.88% and 3.75% on average F1 compared to MCMN and PRM, respectively. In the 5-way 5-shot setting, RoFRC achieves an improvement of 1.68% and 2.07% on average accuracy and F1 respectively, compared to PRM. The results indicate the effectiveness of RoFRC.

Table 1 can also prove that RoFRC has strong ro-

bustness. In the 5-way 1-shot scenario, the PRM accuracy and F1 decrease by 4.61% and 1.76% from 50% to 15% NOTA rate, whereas our approach reduces only by 2.43% and 0.77%, respectively. In the 5-way 5-shot scenario, MCMN's accuracy and F1 decrease by 12.49% and 4.9%, respectively, from 15% to 50% NOTA rate. In contrast, PRM's accuracy and F1 reduce by 3.14% and 1.29%, respectively, from 50% to 15% NOTA rate. The RoFRC only exhibits a slight decrease of 1.4% and 0.31%, respectively. It is worth noting that following the original experimental settings (Gao et al., 2019), all models are trained at a 50% NOTA rate and evaluated at various rates (see §A.2). Therefore, it can be concluded that RoFRC exhibits more robust performances than baseline methods.

The results in Table 2 also demonstrate the superiority of the RoFRC model. Whether in the 1-shot or 5-shot scenario, the average performance of RoFRC surpasses the baseline. Particularly in terms of F1 score, RoFRC significantly outperforms the strong baseline. Moreover, the results of O-Proto and DProto in Table 2 also demonstrate that using a binary classification method to detect NOTA instances after the classification of the prototypical network is not suitable for a small-scale dataset. This vector distribution-based NOTA detection method cannot effectively learn the features of NOTA instances.

4.3 Online Evaluation

CodaLab (Pavao et al., 2022) is an open-source platform that enables researchers, developers, and data scientists to collaborate. In order to ensure fairness and reproducibility of research work, Gao et al. (2019) upload the FewRel 2.0 test set on the *CodaLab*. As there are no limitations on the experimental conditions (e.g. hyperparameters), all submissions in *CodaLab* chase the optimal performance on the NOTA challenge. We submit the test results of FewRel 2.0 to the *CodaLab* platform using our best model. By using the same evaluation criteria and submitting predictions for the same dataset, we can directly compare our results with those of other submissions. The results presented in Table 3 demonstrate that RoFRC performs optimally in the NOTA challenge, with an average accuracy that exceeds MCMN by 2.33 percentage points. These findings offer strong evidence of RoFRC's effectiveness and robustness.

Methods	**5-way 1-shot**		**5-way 5-shot**	
	15%	50%	15%	50%
Proto(CNN)	60.59	40.00	77.79	61.66
Proto(BERT)	70.02	45.94	83.79	75.21
BERT-Pair	77.67	80.31	84.19	86.06
MNAV	79.06	81.69	85.52	87.74
anonymous1	67.97	74.85	81.94	78.12
anonymous2	79.53	79.99	86.31	81.92
MCMN	88.40	84.56	89.91	85.32
PRM	83.01	83.32	89.30	85.94
RoFRC (Ours)	**89.67**	**87.40**	**91.30**	**89.16**

Table 3: Online evaluation results on the FewRel 2.0 test set, in terms of accuracy (%). All the results are obtained from *CodaLab*, where the NOTA rates are specified as 0.15 and 0.5.

Methods	**5-way 1-shot**		**5-way 5-shot**	
	15%	50%	15%	50%
Grok-2	80.99	80.95	85.28	79.20
Gemini 2.0 Flash	88.37	79.50	89.12	75.05
GPT-4o	80.75	80.86	83.93	76.54
DeepSeek V3	85.28	82.95	88.92	81.95
Claude 3.5 Sonnet	**88.92**	82.61	**91.82**	82.30
RoFRC (Ours)	84.64	**87.07**	87.38	**88.78**

Table 4: Comparison of accuracy (%) on the FewRel dataset with popular large language models, evaluated at NOTA rates of 0.15 and 0.5.

4.4 Compare with Large Language Models

Furthermore, we assess the performance of the popular large language models (LLMs) in the NOTA challenge. Recent works (Agrawal et al., 2022; Jeblick et al., 2022; Zhang et al., 2023) have shown that large-scale pre-trained language models, including GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022) are capable of performing well in numerous downstream tasks without parameter tuning. Building on these findings, we conducted experiments on the FewRel dataset to assess the capabilities of several popular LLMs, including Grok-2, Gemini 2.0 Flash, GPT-4o (Achiam et al., 2023), DeepSeek V3, and Claude 3.5 Sonnet.

As shown in Table 4, RoFRC outperforms all evaluated LLMs when the NOTA rate is set to 0.5. However, at a lower NOTA rate of 0.15, it performs less competitively, trailing behind models such as Claude 3.5 Sonnet and Gemini 2.0 Flash. The following analysis explores the underlying reasons for these observations.

At a NOTA rate of 0.15, the primary challenge lies in distinguishing between known classes. Under these conditions, LLMs demonstrate superior performance, likely due to their extensive pretrain-

Methods	5-way 1-shot						5-way 5-shot					
	15%		30%		50%		15%		30%		50%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RoFRC ([CLS])	75.46	74.35	77.51	75.82	81.30	76.15	80.34	79.39	81.71	80.69	84.12	80.56
Relation-level decision	67.03	65.43	71.23	67.76	78.87	69.70	65.82	64.20	70.19	66.52	78.16	68.62
RoFRC without $\mathcal{L}_{agree}$	81.25	80.44	83.00	81.99	86.00	82.42	86.73	86.08	87.51	87.05	88.99	86.91
RoFRC ($\mathcal{L}_{agree}$ only non-NOTA)	83.45	82.69	84.78	84.06	86.89	84.06	86.81	86.18	87.58	87.13	**89.00**	86.97
RoFRC(Ours)	**84.64**	**83.92**	**85.59**	**85.05**	**87.07**	**84.69**	**87.38**	**86.75**	**87.76**	**87.44**	88.78	**87.06**

Table 5: Ablation study results of RoFRC, in terms of accuracy (%) and F1 (%), on the FewRel dataset. The reported results are the average and standard deviation of five runs.

ing on large-scale corpora, which enhances their ability to generalize across diverse relation types. Furthermore, in low-NOTA scenarios, classification accuracy is largely dependent on fine-grained feature discrimination. LLMs may be particularly effective in capturing subtle semantic distinctions between relation types, leading to higher accuracy.

In contrast, when the NOTA rate increases to 0.5, the classification task becomes significantly more challenging, as a substantial proportion of test samples do not belong to any predefined category. In this setting, RoFRC outperforms all LLMs, highlighting its stronger capability in rejecting out-of-distribution samples. This advantage can be attributed to several factors. First, RoFRC employs specialized loss functions that enhance its robustness against unknown samples. Second, LLMs often exhibit overconfidence in classification tasks, which increases the likelihood of misclassifying NOTA samples into known categories.

In summary, the experimental results suggest that while LLMs excel in scenarios requiring fine-grained classification with a low NOTA rate, they exhibit limitations in high-NOTA settings. RoFRC, on the other hand, demonstrates robustness in handling NOTA samples.

4.5 Ablation Study

To explore the impact of individual modules, we perform an ablation study. The results are reported in Table 5. Firstly, compared with ROFRC ([CLS]), it is clearly evident that the [CLS] token embedding as the token's vector representation of the relation description and instance results in a significant decrease in performance. This establishes the effectiveness of the embedding technique we used and provides the best performance for downstream relation classification. Secondly, it is observed that only using the relation-level decision causes significant performance drops. This suggests that blending two decisions is effective in the proposed method, thereby ignoring the role of support instances is inappropriate. Thirdly, after removing $\mathcal{L}_{agree}$, the model's performance consistently degrades compared to RoFRC. Especially, in the 5-way 1-shot setting, both the average accuracy and F1 decrease significantly by 2.35% and 2.93%, respectively. This validates the positive impact of $\mathcal{L}_{agree}$ in helping the model boost performance and ensuring the consistency of two decisions.

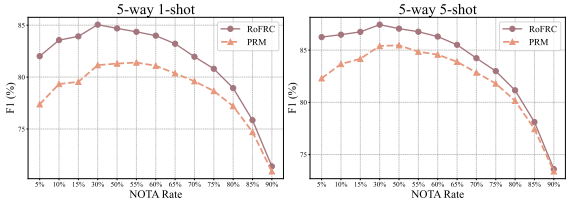

Figure 3: NOTA rates analysis. The robustness of our method at various NOTA rates is demonstrated by comparing it with the SOTA PRM. The abscissa represents different NOTA rates.

An important question arises when considering $\mathcal{L}_{agree}$: *should we include all query instances (non-NOTA and NOTA) when computing $\mathcal{L}_{agree}$?* In order to explore this question, we compare the results to RoFRC using $\mathcal{L}_{agree}$ with only non-NOTA instances. The experimental results demonstrate that exclusively using non-NOTA instances for $\mathcal{L}_{agree}$ loss results in a decline in model performance, particularly at the 15% NOTA rate. One possible reason is that for NOTA instances, increasing $\mathcal{L}_{agree}$ may result in an elevated predicted probability of a specific non-NOTA category that may not belong to its correct class. Balancing the consistency of the two decisions by utilizing NOTA instances is an effective strategy to enhance the prediction accuracy of non-NOTA instances.

4.6 NOTA Rate Analysis

To demonstrate the robustness of our approach, we evaluate the F1 performance of our model at various NOTA rates. Specifically, we add extra eight NOTA rates larger than 50% and extra two NOTA rates smaller than 15% when sampling episodes, as illustrated in Figure 3. We adopt the episode sampling approach of Gao et al. (2019), with the difference being that we sample a greater or smaller number of NOTA query instances within each episode. As an illustration, suppose we sample 5 non-NOTA instances and 7 NOTA instances. The resulting NOTA rate would be calculated as 7 divided by the sum of the number of non-NOTA instances and NOTA instances. $(7/(5+7)) * 100\% \approx 60\%$.

Our comparative analysis between RoFRC and PRM reveals that RoFRC outperforms PRM with higher F1 in all NOTA rates, as indicated in Figure. 3 . Therefore, we can conclude that RoFRC is more robust and adaptable to different NOTA rates.

5 Conclusion

This paper introduces the RoFRC architecture, a novel approach specifically designed to address the challenge of NOTA in the few-shot relation classification task. The primary objective is to enhance the model's performance and robustness across varying rates of NOTA. The proposed architecture includes a decision blender that efficiently performs few-shot relation classification and adapts to the NOTA scenario by merging instance-level and relation-level decisions. To ensure consistency, an agreement loss function is introduced to weigh the agreement between instance-level and relation-level decisions. Experimental results on a popular dataset FewRel confirm the superior performance of our RoFRC architecture.

Limitations

This study addresses the few-shot relation classification task, which includes NOTA instances, and evaluates the performance of the proposed RoFRC architecture. Our results indicate that the RoFRC architecture is efficient and robust, but a limitation of RoFRC is its dependency on relation descriptions for every relation in the support set, without which the model's functionality is compromised. However, data lacking relation descriptions can be addressed by utilizing a more expansive language model to automatically generate the necessary relation descriptions, and we have a test in Appendix B.4. In the future, our research will focus on integrating a relation description generation module to adapt to a wide range of real-world data.

Acknowledgements

We sincerely thank all the anonymous reviewers for providing valuable feedback. This work is supported by the National Natural Science Foundation of China (No.62406151).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205. 12689*.

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.

Philipp Borchert, Jochen De Weerdt, and Marie Francine Moens. 2024. Efficient information extraction in few-shot relation classification through contrastive representation learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 638–646.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Ye Dong, Rong Yang, Junbao Liu, and Xizhong Qin. 2024. Few-shot relation extraction through prompt with relation information and multi-level contrastive learning. *IEEE Access*.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging Few-Shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for Few-Shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale supervised Few-Shot relation classification dataset with State-of-the-Art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4803–4809.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Others. 2022. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212. 14882*.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705. 03645*.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. 2022a. Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022b. Learn from relation information: Towards prototype representation rectification for Few-Shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1822–1831, Seattle, United States. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022c. A simple yet effective relation information guided approach for Few-Shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and Others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. 2022. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202. 04414*.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. CodaLab competitions: An open source platform to organize scientific challenges. *Technical report*.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 7867–7876.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems (NeurIPS)*, 30.

Hui Sun and Rongxin Chen. 2025. Enhancing the prototype network with local-to-global optimization for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2668–2677.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-Domain detection for Low-Resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572.

Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810. 03548*.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95.

Yuxia Wang, Karin Verspoor, and Timothy Baldwin. 2020. Learning from unlabelled data for clinical semantic textual similarity. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233, Online. Association for Computational Linguistics.

Jianfeng Wu, Mengting Hu, Yike Wu, Bingzhe Wu, Yalan Xie, Mingming Liu, and Renhong Cheng. 2023. Density-Aware prototypical network for Few-Shot relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2477–2489, Singapore. Association for Computational Linguistics.

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*, CIKM '20, pages 2273–2276, New York, NY, USA. Association for Computing Machinery.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How would stance detection techniques evolve after the launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Ningyu Zhang, Shumin Deng, Zhanling Sun, Xi Chen, Wei Zhang, and Huajun Chen. 2018. Attention-Based capsule networks with dynamic routing for relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (ACL)*, pages 986–992.

A Reproducibility

A.1 Dataset

FewRel. We choose a widely used public dataset, i.e. FewRel 2.0 (Gao et al., 2019) to evaluate the proposed approach. Its first version is the FewRel dataset (Han et al., 2018). Later, Gao et al. (2019) update it to FewRel 2.0, which is constructed from Wikipedia and consists of 100 relations, each with 700 labeled instances. We conduct our experiments using the same data splits as the official FewRel benchmark. Specifically, the dataset is partitioned into 64 classes for training, 16 for validation, and 20 for testing. To ensure the fairness and reproducibility of experiments, we utilize identical episode sampling and evaluate our models on the standard test set.

NYT-25. The NYT-25 dataset is derived from the New York Times corpus, and the original data can be found on the FewRel website. The NYT-25 dataset provides 25 relations, each containing 100 instances. However, the dataset has not been split into training, validation, and test sets. Referring to the dataset division method of Qu et al. (2020), and in order to meet the experimental setting of the NOTA challenge, we randomly sample 10 relations for training, 6 relations for validation, and 9 relations for testing.

A.2 Implementation Details

We utilize the same set of hyper-parameters as Gao et al. (2019) for a fair and equal comparison. Our work and all baselines employ the identical encoder, **BERT-base-uncased** (Devlin et al., 2019). We set the batch size to 2 for training, implying that two episodes are fed concurrently into the model per batch. The model's training comprises 30,000 batches, where each epoch contains 1,000 batches. After each epoch, we evaluate 1,000 batches of the validation set to search for the best model. In addition, we implement the early stopping strategy, which halts the training process if the model's performance fails to improve after 6 consecutive epochs on the validation set. For testing, we test a total of 10,000 batches on the test set. Following Gao et al. (2019), we conduct testing at NOTA rates of 15%, 30%, and 50%, after training with a 50% NOTA rate. All the results reported using the FewRel dataset are the average of five fixed seed runs. In Eq (16), we set the hyper-parameters λ to 1e-5. Following Gao et al. (2019), the remaining hyperparameters are detailed as follows:

- Learning rate: 2e-5.
- Weight decay: 1e-5.
- Warmup steps: 300.
- Gradient accumulation: 1.
- Instance max length: 128.
- Relation description max length: 128.
- Hidden size: 768.
- Seeds: [5, 10, 15, 20, 25].

.

A.3 Evaluation Metrics

Two key evaluation metrics are employed in our work: accuracy and macro F1. The basic implementation of accuracy is based on the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Methods	Number of parameters
O-Proto	109482240
MNAV	109497600
BERT-Pair	109483778
MCMN	109482240
PRM	109494532
RoFRC(Ours)	109482240

Table 6: The number of parameters for each model.

where TP are true positives, FP are false positives, TN are true negatives, and FN are false negatives. The F1 is defined as follows:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \times P \times R}{P + R}$$

Macro F1 calculates the unweighted mean of F1 scores for each class.

A.4 Number of Parameters per Method

Table 6 shows the number of parameters for each model.

A.5 Time Requirement

Methods	Training	Testing		
		15%	30%	50%
O-Proto	17678	2840	2832	3187
Bert-Pair	21211	3389	3738	5244
MNAV	6709	1562	1422	1346
MCMN	3966	9108	9211	9123
PRM	25522	2956	2840	2932
RoFRC	28823	3367	3032	3592

Table 7: The training and testing times (in seconds) for each model in a 5-way 1-shot scenario.

Table 7 records the training and testing time requirements for all models in the 5-way 1-shot scenario.

A.6 Prompt for Large Language Models

Table 12 presents four examples of the LLMs accomplishing a 5-way 1-shot relation classification task. We devise a prompt for LLMs to comprehend the FSRC task and to emphasize the NOTA definition. To assist LLMs in classifying instances in the query set, the prompt includes the relation name and description for each category when inputting the support set.

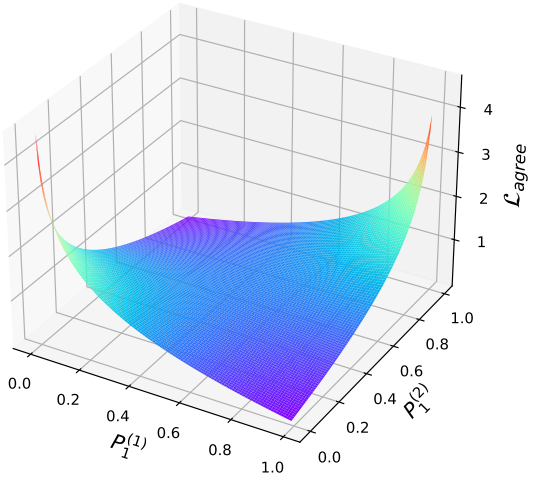

Figure 4: Agreement loss visualization. As $\mathcal{L}_{agree}$ increases, the instance-level decision probability $p_1^{(1)}$ and the relation-level decision probability $p_1^{(2)}$ tend to be the same.

B Overall Results

B.1 Visualization

In order to further demonstrate the effectiveness of the agreement loss in Eq. (15), we visualize the function graph of $\mathcal{L}_{agree}$ in Figure 4. The figure illustrates that as we increase $\mathcal{L}_{agree}$, the instance-level decision probability $p_1^{(1)}$ and the relation-level probability $p_1^{(2)}$ progressively approach each other. Additionally, influenced by the cross-entropy loss $\mathcal{L}_c$ represented in Eq. (11), the blending decision probability of a specific non-NOTA class gradually increases. As a result, the probabilities of $p_1^{(1)}$ and $p_1^{(2)}$ ultimately reach a convergence point of 1. This process ensures consistency between the instance-level decision and the relation-level decision, consequently enhancing the performance of the model.

B.2 Hyperparameter Analysis

The results obtained for different values of the hyperparameter λ in Eq. (16) are illustrated in Table 8. The experimental results indicate that by setting the λ to 1e-5, the optimal performance is achieved under the 5-way 1-shot and 5-way 5-shot settings. Furthermore, the experimental results demonstrate that the overall performance of the model remains relatively stable even when modifying the hyperparameter. Particularly, under the 5-way 5-shot setting, the accuracy and F1 score demonstrate a significant numerical proximity. For comparison, the results of PRM are also presented in Table 8. It

Methods	5-way 1-shot						5-way 5-shot					
	15%		30%		50%		15%		30%		50%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
PRM	80.38	79.54	82.10	81.15	84.99	81.30	84.86	84.17	86.01	85.40	88.00	85.46
RoFRC(λ=1e-3)	84.28	83.54	85.28	84.71	86.84	84.40	85.66	84.99	86.67	86.08	88.58	86.15
RoFRC(λ=1e-4)	81.43	80.62	83.16	82.16	86.00	82.48	86.85	86.21	87.56	87.11	**88.94**	86.91
RoFRC(Ours, λ=1e-5)	**84.64**	**83.92**	**85.59**	**85.05**	**87.07**	**84.69**	**87.38**	**86.75**	**87.76**	**87.44**	88.78	**87.06**
RoFRC(λ=1e-6)	81.94	81.16	83.65	82.70	86.59	83.19	86.86	86.21	87.44	87.04	88.75	86.79

Table 8: Hyperparameter analysis results of RoFRC, in terms of accuracy (%) and F1 (%), on the FewRel dataset. The hyperparameter robustness of RoFRC is demonstrated by comparing the experimental results of PRM.

is apparent that RoFRC consistently outperforms PRM in terms of accuracy and F1 score, irrespective of the value assigned to the hyperparameter λ. In summary, RoFRC exhibits robustness towards hyperparameter fluctuations and shows minimal impact on its performance.

B.3 Performance Gap Analysis

MNAV	15%	30%	50%	Average
1-shot (Non-NOTA)	83.95	84.05	83.96	83.99
1-shot (NOTA)	75.95	75.78	75.91	75.88
5-shot (Non-NOTA)	90.21	90.28	90.22	90.24
5-shot (NOTA)	79.99	79.78	79.79	79.86
PRM	15%	30%	50%	Average
1-shot (Non-NOTA)	80.52	80.66	80.66	80.61
1-shot (NOTA)	91.88	91.72	91.82	91.81
5-shot (Non-NOTA)	85.13	85.14	85.21	85.16
5-shot (NOTA)	92.60	92.64	92.60	92.61
RoFRC(Ours)	15%	30%	50%	Average
1-shot (Non-NOTA)	85.85	86.08	86.04	85.99
1-shot (NOTA)	90.59	90.56	90.61	90.59
5-shot (Non-NOTA)	88.68	88.47	88.60	88.58
5-shot (NOTA)	90.92	90.92	90.93	90.92

Table 9: The accuracy(%) of the model in classifying non-NOTA instances and NOTA instances under different NOTA rates.

In order to analyze the performance gap of the model in recognizing non-NOTA instances and NOTA instances, we test the accuracy of the model based on prototypical network in classifying non-NOTA instances and NOTA instances under the experimental settings described in Section A.2. The results of the test set are shown in Table 9. Firstly, RoFRC demonstrates excellent performance in classifying NOTA instances in both 1-shot and 5-shot scenarios. Secondly, compared with PRM and MNAV, the performance gap in classifying RoFRC between non-NOTA instances and NOTA instances is relatively small in both the 1-shot and 5-shot scenarios. Especially in the 5-shot scenario, RoFRC exhibits similar accuracy in classifying both types of instances. We use simple calculations to prove this conclusion. In the 15% NOTA rate scenario, the average performance of PRM is $(85.15\% \times 5 + 92.61\% \times 1)/6 \approx 86.40\%$, while the average performance of RoFRC is $(88.58\% \times 5 + 90.92\% \times 1)/6 \approx 88.97\%$. In the 50% NOTA rate scenario, the average performance of PRM is $(85.15\% \times 5 + 92.61\% \times 5)/10 \approx 88.89\%$, and the average performance of RoFRC is $(88.58\% \times 5 + 90.92\% \times 5)/10 \approx 89.75\%$. Obviously, the changes in RoFRC are smaller. Therefore, the performance gap between the model in classifying non-NOTA instances and identifying NOTA instances determines its robustness. Thirdly, in the 1-shot scenario, due to only having one support instance for each category, the classification ability of each model for non-NOTA instances decreases, and there is a certain gap compared with the classification ability of NOTA instances. In summary, RoFRC demonstrates stronger robustness in classifying non-NOTA and NOTA instances.

B.4 Relation Descriptions' Quality Analysis

Methods	5-way 1-shot					
	15%		30%		50%	
	Acc	F1	Acc	F1	Acc	F1
RoFRC	**84.64**	**83.92**	**85.59**	**85.05**	87.07	**84.69**
RoFRC(change)	84.19	83.49	85.34	84.71	**87.29**	84.66

Table 10: Study on relation descriptions' quality.

We regenerated the relation descriptions in the FewRel dataset using ChatGPT and conducted new experiments under the 5-way 1-shot scenario. The experimental results are presented in the Table 10. As shown, our model still maintains strong performance, demonstrating its robustness to variations in the quality of relation descriptions.

B.5 Traditional FSRC Performance Analysis

Methods	5-way 1-shot		5-way 5-shot	
	Acc	F1	Acc	F1
BERT-Pair	88.68	87.93	94.43	94.04
MNAV	83.98	82.83	93.81	93.43
PRM	93.92	93.57	95.66	95.39
MultiRep	94.18	-	96.29	-
RelPromptCL	94.71	-	**97.38**	-
LoToG	**95.28**	-	96.71	-
RoFRC (Ours)	94.44	**94.05**	96.46	**96.23**

Table 11: Traditional few-shot relation classification results on the FewRel test set.

Some existing works only design models for traditional few-shot relation classification tasks (i.e., the query set does not include NOTA instances), and they do not provide a solution for the NOTA challenge. We have also tested our method on traditional few-shot relation classification tasks on FewRel dataset. We compare our model with the FSRC models, including MultiRep (Borchert et al., 2024), RelPromptCL (Dong et al., 2024), LoToG (Sun and Chen, 2025). The experimental results in Table 11 show that our proposed RoFRC is also competitive.

Question: The purpose of the relation classification task is to identify the relation between two entities (one is the head entity and the other is the tail entity) in a given sentence. For example, the following sentence "California is a state in the United States". Through the derivation of the whole sentence, the relation between the head entity "California" and the tail entity the "United States" is "contains". While the few-shot relation classification task consists of a support set and a query set. The support set generally contains N relations, and each relation contains K instances. This setup is called N-way K-shot. We need to predict the relation between two entities in each sentence given in the query set based on the support set. Now I need you to complete a few-shot relation classification task, which is set to 5-way 1-shot. The five categories in the support set are numbered from 0 to 4, respectively. Five sentences are given after each category. The head entity of each sentence is surrounded by two special characters [head] and [/head], and the tail entity is surrounded by two special characters [tail] and [/tail]. You need to predict the relationship between two entities for each sentence in the query set from the support set. That is, if the relation of a query instance is similar to that of all instances of category 0 in the support set, then the category predicted by the query instance is category 0. Notice! If the relation between the two entities of a sentence in the query set cannot be referenced in the support set, the relation between the two entities of the sentence is called NOTA (None-of-the-above). When the NOTA relation is encountered, the category number of the predicted result is 5. Warning! You just need to return the class number for each instance predicted in the query set. For example, the relation of the first instance in the query set is relation 0, and the second instance is relation 5 (NOTA). Then your reply is: [0, 5]. If you get it, then I'll give you support sets and query sets next.

LLMs: Yes, I understand the task. Please provide me with the support set and query set.

Question: **The following is the support set (5-way 1-shot. Instances in each category are separated by a special character [sep].)**: Name of category 0: chairperson. Description of category 0: presiding member of an organization, group or body. Instances in category 0: Rowley was elected leader of the [head] Communist Party of Canada [/head] in January 2016 by the Party's Central Committee, following the retirement of [tail] Miguel Figueroa [/tail]. [sep] Name of category 1: parent organization. Description of category 1: parent organization of an organization, opposite of subsidiaries (P355). Instances in category 1: In 1932 South Georgia State College emerged as one of the original 26 institutions of the [tail] University System of Georgia [/tail] which traces its origins to the chartering of the [head] University of Georgia [/head] in 1785. [sep] Name of category 2: family. Description of category 2: family, including dynasty and nobility houses. Not family name (use P734 for family name). Instances in category 2: Matthias and Frederick III / V had been rivals stretching back to Matthias' succession as King of Hungary in 1458 after the early death of Frederick's [tail] Habsburg [/tail] cousin King [head] Ladislaus the Posthumous [/head]. [sep] Name of category 3: connecting line. Description of category 3: railway or public transport line(s) subject is directly connected to; use as a qualifier to P197. Instances in category 3: Seibuen Station is the terminus of the [tail] Seibuen Line [/tail], a 2.4 kilometer spur line from [head] Higashi - Murayama Station [/head]. [sep] Name of category 4: parent taxon. Description of category 4: closest parent taxon of the taxon in question. Instances in category 4: The orders Raphidioptera and [head] Neuroptera [/head] are grouped together as [tail] Neuropterida [/tail]. [sep] Please give me the predicted class number for each instance in the query set. For example, if the query set contains two instances, and you make a prediction based on the support set, the first instance has category number 0, and the second instance has category number 5 (NOTA), then you reply to me [0, 5]. **The following is the query set (The query set contains one instance.)**: It is part of the [tail] Epsom and St Helier University Hospitals NHS Trust [/tail] along with the nearby [head] St Helier Hospital [/head] and Sutton Hospital.

LLMs: The predicted class number for the given instance in the query set is 1.

Table 12: An example of using LLMs to complete the FSRC task in the experiment.