CLLMFS: A Contrastive Learning enhanced Large Language Model Framework for Few-Shot Named Entity Recognition

Yafeng Zhang[a,*,1], **Zilan Yu**[b,**,2], **Yuang Huang**[a] **and Jing Tang**[d]

[a]iFLYTEK Co., Ltd.
[b]Tsinghua University
[c]Huazhong University of Science and Technology
ORCID (Yafeng Zhang): https://orcid.org/0000-0001-5619-1721, ORCID (Zilan Yu): https://orcid.org/0000-0002-9460-0984, ORCID (Yuang Huang): https://orcid.org/0009-0004-9084-6807, ORCID (Jing Tang): https://orcid.org/0000-0002-9430-9660

Abstract. Few-shot Named Entity Recognition (NER), the task of identifying named entities with only a limited amount of labeled data, has gained increasing significance in natural language processing. While existing methodologies have shown some effectiveness, such as enriching label semantics through various prompting modes or employing metric learning techniques, their performance exhibits limited robustness across diverse domains due to the lack of rich knowledge in their pre-trained models. To address this issue, we propose CLLMFS, a Contrastive Learning enhanced Large Language Model (LLM) Framework for Few-Shot Named Entity Recognition, achieving promising results with limited training data. Considering the impact of LLM's internal representations on downstream tasks, CLLMFS integrates Low-Rank Adaptation (LoRA) and contrastive learning mechanisms specifically tailored for few-shot NER. By enhancing the model's internal representations, CLLMFS effectively improves both entity boundary awareness ability and entity recognition accuracy. Our method has achieved state-of-the-art performance improvements on F1-score ranging from 2.58% to 97.74% over existing best-performing methods across several recognized benchmarks. Furthermore, through cross-domain NER experiments conducted on multiple datasets, we have further validated the robust generalization capability of our method. Our code will be released in the near future.

1 Introduction

Named Entity Recognition (NER) is pivotal for identifying and categorizing named entities within unstructured text across various domains, such as Location [23], Private Health Information [21] and Event [32]. However, developing accurate NER models demands substantial amounts of domain-specific annotated data, which are often scarce and costly to procure [14]. This has led to the demand for Few-Shot Named Entity Recognition (FS-NER), which aims to learn from limited labeled examples to address entity tagging challenges under low-resource conditions [6].

Early FS-NER methods often employ neural networks with conventional supervised learning, which may lead to overfitting due to the large number of parameters to optimize [3]. To mitigate this, cross-domain NER approaches have been employed, where models learn semantic features from base classes and adapt them to novel classes [16]. Despite this, these methods may still exhibit suboptimal generalization in novel domains [9]. To address these limitations, contrastive learning has been introduced, utilizing Gaussian distributions to optimize the distributional distance between tokens in sentences [7].

With the rapid development of Large Language Models (LLMs), models like GPT-3 demonstrate few-shot capabilities through prompt-based construction, achieving satisfactory results [30]. LLAMA 2 [25] emerges as a superior choice in low-resource settings due to its accessibility and adeptness across various natural language processing (NLP) tasks. However, deploying of LLMs such as ChatGLM, GPT-3, ChatGPT, GPT-4, LLAMA, and LLAMA 2 [10, 4, 17, 5, 24, 26] for FS-NER poses challenges, given their extensive parameter sizes and the need for substantial amounts of high-quality supervised fine-tuning (SFT) data, leading to high costs in training and data acquisition. To address this, Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation (LoRA) [13] have been proposed to enhance model performance on new tasks while minimizing fine-tuning parameters and computational complexity.

In this research, we present an innovative method, CLLMFS, to tackle the NER task under low-resource conditions. Our approach leverages LLMs to effectively address the challenge of limited labeled data by exploiting their pre-trained knowledge. We fine-tune LLMs using supervised learning to adapt

[*] Corresponding Author. Email: yfzhang40@iflytek.com.
[**] Corresponding Author. Email: yuzl22@mails.tsinghua.edu.cn.
[1] Equal contribution.
[2] Equal contribution.

them to our specific NER task, resulting in improved performance compared to recent benchmarks. To further reduce the trainable parameters, we employ LoRA techniques, enabling effective fine-tuning with limited training samples. Additionally, we introduce contrastive learning to our framework, enriching the LLM-based method for few-shot NER tasks. This framework significantly improves the boundary awareness of the LLM and enhances its ability to accurately extract named entities by refining internal embedding representations. Furthermore, we enhance the model's robustness by introducing noise to construct positive example pairs during training. Our approach achieves state-of-the-art results across multiple datasets, demonstrating its effectiveness and versatility in handling NER tasks under low-resource settings.

In summary, our contributions are as follows:

- We advance the use of LLMs for few-shot NER by integrating them with LoRA for supervised fine-tuning, achieving the-state-of-art performance across multiple datasets with limited labeled data.
- We propose a framework that incorporates contrastive learning to improve the boundary awareness and accuracy of entity extraction, enhancing model robustness by constructing positive example pairs with noised embedding.
- Our approach showcases robust transfer capabilities, significantly enhancing the F1-score, ranging from 2.58% to 97.74% in the INTRA setting, and from 44.36% to 160.00% in the INTER setting, surpassing state-of-the-art methods across various datasets.

2 Related Works

2.1 *Few-Shot NER*

Few-shot learning (FS-NER) enhances model performance with limited labeled data. Data-enhancement methods augment small labeled datasets with additional data sources, but unreliable examples can affect precision [33]. Manner uses a Variational Autoencoder for an external memory module, but faces challenges in memory optimization and cross-domain generalization [11]. CONTaiNER employs contrastive learning to optimize token distribution, improving adaptability to new domains [7], but large source-target domain divergence can be problematic.

2.2 *Meta Learning*

Meta Learning offers new approaches for few-shot learning. Metric-based methods like Matching Networks [27] and Prototypical Networks [19] calculate similarities to learn prototypical representations for target classes. ProtoBERT uses a precomputable BERT encoder for effective entity prediction [37]. ProML introduces multiple prompt schemas with weighted averages for enriched label semantics [6], achieving promising results across various settings. However, generated prototypes may lack precision, due to limited labeled data for various entity types in the support set.

2.3 *In-context learning*

Large-scale pre-trained LLMs, like GPT-3 [4], have advanced in-context learning, applied in tasks like question answering and NER without additional training data [2]. Recent NLP research explores prompt-based methods for FS-NER, relying on prompts to predict labels. However, these methods primarily rely on prompts to predict labels using classification heads, rather than employing data-enhancement or metric learning techniques. Prompt-based NER uses language models to generate entity predictions based on context and instructions. However, these methods face limitations in prompt quality and design.

3 Methodology

The internal representations of language models play a pivotal role in shaping the performance of downstream tasks. In this paper, we introduce a novel model, CLLMFS, based on large pre-trained language models. As depicted in Figure 1, our model undergoes supervised fine-tuning in source domains under the *N-way K-shot* scenario, enabling it to adapt to target domains effectively. Ultimately, our model integrates LoRA and contrastive learning loss techniques, specifically customized for the NER task.

3.1 *Task definition*

Given a sequence of n tokens $\{x_1, x_2, \ldots, x_n\}$ and corresponding tag labels $\{y_1, y_2, \ldots, y_n\}$, the primary objective of NER is to associate each token x_i with its corresponding tag label y_i. In Few-shot NER, a model undergoes training in a low-resource source domain with a tag-set denoted as $\{C_i^s\}$. Subsequently, it is tested in a target domain that employs a distinct tag-set, denoted as $\{C_j^d\}$, where i and j represent indices for different tags. Since $\{C_i^s\} \cap \{C_j^d\} = \varnothing$, the model faces the formidable challenge of generalizing to previously unseen test tags. In an *N-way K-shot* scenario, the source domain comprises N distinct entity types, denoted as $|\{C_j^s)\}| = N$. For each entity type, there are K examples in the support set. This setup means that the model is trained with K labeled examples for each of the N types, enabling it to learn and generalize from a limited number of examples.

3.2 *LLM for entity extraction*

Our method for entity extraction tasks is based on LLAMA 2, referred to as LLM. Figure 1 illustrates the pivotal components of our model. We utilize the 7-billion parameter model of LLAMA 2, balancing effectiveness and inference speed. Our approach achieves excellent results in entity extraction tasks with only a small amount of training samples.

LLAMA 2's architecture closely resembles the standard Transformer Decoder, primarily consisting of 32 Transformer Blocks. Each block includes the following core components:

- **RMSNorm** [36]: Normalizes the activation outputs of network layers, ensuring uniform scaling, accelerating training, and enhancing model stability.
- **SwiGLU** [18]: Adds non-linearity to the model by transforming input values through the Swish activation function.
- **RoPE** [22]: A novel positional encoding strategy that encodes positional information through rotation operations.

- **GQA** [1]: Divides query heads into G groups, with each head maintaining its own query parameters and each group sharing a key and value matrix, simplifying calculations and improving the efficiency of attention computation in large models.

By leveraging these components, our method effectively extracts entities with high accuracy and efficiency.

3.3 Model Training

3.3.1 Supervised fine-tuning with LoRA

Supervised fine-tuning (SFT) refers to the process of adjusting a pre-trained LLM using labeled data to better adapt it to a specific task. During SFT, weights of the model are adjusted based on the discrepancies with the true labels, aiming to enhance precision and task adaptation.

Each sample in SFT typically consists of three parts: instruction (i.e., prompt), input, and output. For instance, for the entity type "Person" (other entity types are provided in the Appendix A):

```
 { "instruction": "Please extract the Person in
the sentence given below, the entity of person
refers to the entity that represents the identity or
role of a specific person in the input sentence.",
"input": "True , but I imagine it would be a lot
lower and as I pointed out to Andrew Little would
be cheaper than [ eliminating fees .", "output":
"<im_start> I can extract entities for you, the
extracted entities are <<< Andrew Little >>>
<im_end>" }
```

The design concept behind constructing SFT data involves using the instruction to define the entity extraction task and the types of entities to be extracted, guiding the LLM to efficiently perform entity extraction tasks. The input represents the original input of the user, containing the sentences from which entities are to be extracted. For example, Our entity type token in `instruction` is "Person", and the actual entity in the `input` sentence is "Andrew Little". The output denotes the output results of the model, with the extracted entities surrounded by start (<<<) and end (>>>) symbols. Additionally, we intentionally devised a specific format for the output of the LLM, starting with `<im_start>` and ending with `<im_end>`.

Considering the limited amount of the data generated by SFT, full model fine-tuning is not feasible and the issue of overfitting is also serious. We adopt Low-Rank Adaptation (LoRA) [13] to address these problems. LoRA assumes that weight updates during the adaptation process also have a lower 'intrinsic rank'. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, we restrict its update through a low-rank decomposition as follows:

$$W_0 + \Delta W = W_0 + BA \tag{1}$$

where $B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$, and the rank $r << \min(d, k)$. Throughout training, W_0 remains frozen and does not undergo gradient updates, while A and B include trainable parameters. We employ a random Gaussian initialization for A and set B to zero, ensuring that $\Delta W = BA$ is zero at the start of training.

In other words, during the fine-tuning process, the model initializes with pre-trained parameters W_0 and updates them to $W_0 + \Delta W(\theta)$ by maximizing the conditional language model probability, where $|\theta| << |W_0|$:

$$\max_{\theta} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(P_{W_0+\Delta W(\theta)}(y_t|x, y < t)) \tag{2}$$

where Z denotes the training dataset comprising input sequences x and their corresponding target sequences y, and $|y|$ signifies the length of the target sequence y. $P_{W_0+\Delta W(\theta)}(y_t|x, y < t)$ represents the probability that the model predicts the t-th element y_t of the target sequence, given the input sequence x and the first t elements of the target sequence y. LoRA fine-tune only need a subset of parameters, thereby avoiding issues such as excessive resource consumption caused by full fine-tuning.

In principle, LoRA can be applied to any subset of weight matrices in a neural network, thereby reducing the number of trainable parameters.

3.3.2 Decoding strategies

Decoding strategies play a pivotal role in text generation tasks. In our approach, we maintain a fixed temperature of 0.01 to control the diversity of generated text. Additionally, we employ the top-k sampling method, where only tokens ranking within the top probability threshold (top_p) are considered during the token sampling process. This strategy optimizes the greedy approach by sampling from the top-k tokens, allowing tokens with higher scores or probabilities beyond the top threshold to also have a chance of being selected.

Given the nature of our task, which involves information extraction, we implement a constrained generation approach. This ensures that the generated output is constrained to be a subset of the input, restricting the generated content within predefined boundaries. Furthermore, to prevent the model from endlessly generating content, we introduce a custom stop symbol, denoted as `<im_end>`, marking the end of the generated sequence. This mechanism effectively halts the generation process after the last eight characters, ensuring controlled and targeted generation. Employing a combination of a low-temperature setting and constrained generation proves instrumental in effectively handling few-shot scenarios.

The internal representations of language models have a significant impact on the performance of downstream tasks. In this paper, we employ contrastive learning loss in low-resource entity extraction tasks to enhance the boundary perception ability of the model and improve its effectiveness.

3.4 Contrastive learning

Contrastive learning is a discriminative representation learning method based on the principle of comparison, primarily used for unsupervised (self-supervised) representation learning. The core idea of contrastive learning is to compare samples with positive examples (semantically similar) and negative examples (semantically dissimilar). By designing contrastive losses, it aims to bring representations of semantically similar positive examples closer while pushing representations of semantically dissimilar negative examples further apart. Therefore, the careful selection of positive and negative sample pairs for contrastive learning is crucial.

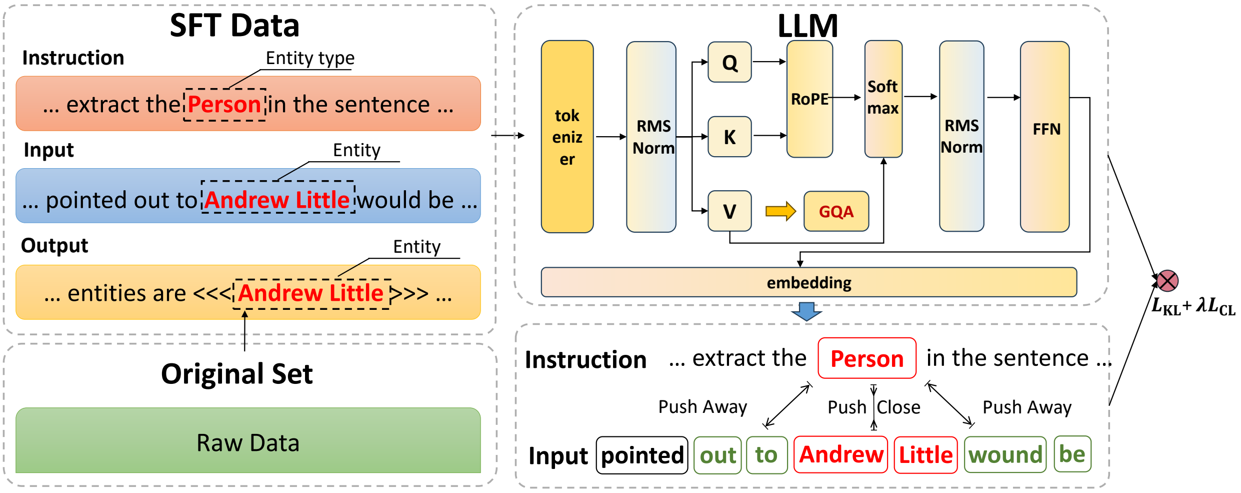

Figure 1. The framework overview CLLMFS. The LLM extracts named entities from carefully designed SFT data using decoding strategies, LoRA fine-tuning leveraging LLM's attention mechanisms such as QKV computations, constructing positive and negative samples for contrastive learning, and creating adversarial embedding samples.

We propose a specific approach to address the challenge of designing positive and negative samples for contrastive learning in low-resource entity extraction tasks. In the constructed SFT data, the entities to be extracted from the input are designated as positive samples. Additionally, the neighboring entities of the target entities serve as negative samples, emphasizing the significance of capturing the entity boundaries accurately. As shown in fig. 1, the embedding of "Person" is proximately aligned with that of "Andrew Little", while being intentionally distanced from the embeddings corresponding to "out to" and "wound be". It is essential to avoid over-extracting or under-extracting words, particularly ensuring against over-extracting.

We define $\mathcal{T}$ as the set of embeddings of entity types within the instructions, and $\mathcal{E}$ as the set of embeddings of entities from input sentences. Therefore, we treat the entity type embeddings within the instructions and the embeddings of the entities to be extracted from input sentences as the positive pairs (*i.e.*,$\{(z^t_{instr}, z^e_{in})|t \in \mathcal{T}, e \in \mathcal{E}\}$). Simultaneously, we establish negative pairs (*i.e.*,$\{(z^t_{instr}, z^n_{in})|t \in \mathcal{T}, n \in \mathcal{E}\}$) by considering the entity type embeddings within the instructions and the neighboring entities of the entities to be extracted from input sentences. Formally, we employ the contrastive loss, InfoNCE [12], to maximize agreement among positive pairs and minimize it among negative pairs:

$$L_{CL} = \sum_{t\in\mathcal{T}, e\in\mathcal{E}} -\log \frac{\exp\left(s(z^t_{instr}, z^e_{in})/\tau\right)}{\sum_{n\in\mathcal{E}, n\neq e} \exp\left(s(z^t_{instr}, z^n_{in}/\tau\right)} \tag{3}$$

where $s(\cdot)$ denotes the similarity between two vectors and is set as the cosine similarity function. τ, referred to as the Temperature parameter in the softmax function, is a hyper-parameter.

By employing this design for positive and negative samples, the distance between the entity type embeddings within the instructions and the embeddings of the entities to be extracted from input sentences is minimized. This adjustment enables the model to prioritize positive entities more effectively during generation. Simultaneously, it increases the distance between the entity type embeddings within the instructions and the neighboring entities of the entities to be extracted from input sentences. This enhances the model's boundary perception ability, resulting in more precise extraction of entity information.

3.5 *Enhancing Representation Uniformity with Adversarial Samples*

The representations generated by contrastive learning are typically regularized, causing them to concentrate within a hypersphere. Alignment and uniformity refer to two essential characteristics of a good representation space: alignment ensures that representations of semantically similar samples are close together, while uniformity ensures that representations of semantically dissimilar samples are evenly distributed across the hypersphere. Enhancing the uniformity of representation distributions can improve the performance of many tasks, such as recommendation systems.

However, previous research has primarily relied on in-batch negative sampling or random negative sampling from the training data. This approach may introduce sampling bias, leading to the inclusion of inappropriate negative examples (such as false negatives or anisotropic representations) in contrastive learning, potentially compromising the alignment and uniformity of the representation space.

To achieve a more uniformly distributed representation space, we focuses on the embedding space and directly introduces noise into the representations. Inspired by Yu et al. [34], we construct adversarial samples through imperceptible perturbations by adding uniformly distributed Gaussian random noise to positive embeddings of entities. While this approach is simple, it can strengthen the positive samples to resist noise, leading to a significant enhancement in the model's robustness against interference. Formally, given a token *i* and its embedding z_i in the *d*-dimensional space, we can implement the following representation-level augmentation:

$$z_i' = z_i + \Delta_i' \tag{4}$$

where Δ_i' is the added noise vectors.

3.6 Model Optimization

To train our model effectively for the low-resource entity extraction task, we employ a combined loss function comprising both cross-entropy loss and contrastive learning loss.

The primary objective of cross-entropy loss in Few-shot NER is to ensure that our model learns to correctly associate each token x_i in the input sequence with its corresponding tag label y_i. This involves minimizing the discrepancy between the predicted tag probabilities and the ground truth labels across the entire sequence. Formally, the cross-entropy loss L_{CE} is computed as follows:

$$L_{CE} = -\sum_{i} \sum_{c \in \{C_j^d\}} y_{i,c} \log(\hat{y}_{i,c}) \tag{5}$$

where $y_{i,c}$ represents the ground truth label for token x_i corresponding to tag c in the target domain, and $\hat{y}_{i,c}$ represents the predicted probability of token x_i belonging to tag c.

By minimizing the cross-entropy loss, our model learns to accurately predict the tags associated with each token in the input sequence, thereby improving its performance in the NER task, especially when dealing with previously unseen tags in the target domain.

In addition to cross-entropy loss, we incorporate contrastive learning loss to further enhance the model's performance. The contrastive loss L_{CL} encourages the model to effectively distinguish between positive pairs (tokens associated with the same entity type) and negative pairs (tokens associated with different entity types).

Due to the presence of 32 Transformer Blocks (i.e., 32 layers of hidden states) in LLAMA 2, we determined the optimal layer for computing the contrastive loss through empirical testing. Configurations using the 10th, 25th, 26th, 27th, and 30th layers were evaluated, and the 26th layer consistently yielded the best performance. This layer selection closely aligns with the 8:2 golden ratio, providing a balance between the lower and higher layers in the model's architecture. Therefore, we compute the contrastive loss at the 26th layer to leverage this optimal configuration.

Finally, we leverage a multi-task training strategy to jointly optimize the cross-entropy loss, and the contrastive learning loss. The overall loss function is:

$$L = L_{CE} + \lambda L_{CL} \tag{6}$$

where λ serves as a hyperparameter to regulate the impact of contrastive learning and is set to 0.001. This choice is made considering that different losses calculate distinct gradients, with the aim of emphasizing the gradient of the main task.

By jointly optimizing cross-entropy loss and contrastive learning loss, our model learns to effectively classify entities while also capturing semantically meaningful representations, thus enhancing its overall performance in the low-resource entity extraction task.

4 Experiments

4.1 Dataset Description

To assess the effectiveness of our method, we utilize 5 datasets spanning various domains: WNUT'17 [3], GUM [4], I2B2 [5], OntoNotes [6], and Conll2003 [7] which are publicly available and have been used in existing research [15, 28, 29, 31] to showcase diversity in terms of domain, scale, and sparsity.

- **WNUT'17** [8] is a collection of noisy user-generated text from social media platforms. This dataset contains annotations for 6 entity types, including 'corporation', 'creative-work', 'group', 'location', 'person', and 'product'.
- **GUM** [35] stands as a versatile, open-source multilayer resource, encompassing a spectrum of twelve text genres including narratives, interviews, news, instructions, and academic writing. It covers 11 entity types such as time, object, quantity, organization and other entities.
- **I2B2** [20] is annotated for Protected Health Information (PHI) and disease Risk Factors, serves as a critical resource within the medical domain. We specially focus on 6 entity recognition of PHI, like 'Patient ID', 'Hospital Location', 'Visit Date', 'Patient Profession', and 'Profession Contact'.
- **OntoNotes** [32] is a large-scale, multilingual corpus that is collected from news, conversational telephone speech, weblogs and broadcast. This paper focuses on 18 entity types, including 'Geopolitical Entity', 'Organization', 'Person', 'Location', 'Money', 'Facility', 'Date', 'Ordinal', 'Quantity', 'Time', 'Nationalities, Religious or Political Groups', 'Cardinal', 'Percent', 'Event', 'Work of Art', 'Language', 'Law', and 'Product'.
- **CoNLL'03** [23] is also a benchmark dataset that focuses on 4 types of entities: persons, locations, organizations, and miscellaneous entities that do not belong to the previous three categories.

All of the above datasets use the N-way and 5-shot setting for training. For a fair comparison on those datasets, we split long sentences in some datasets into multiple shorter sentences to accommodate the input token limit of LLM, thus facilitating the extraction of text information by CLLMFS. We conducted tests on the WNUT'17, GUM, I2B2, OntoNotes, and CoNLL'03 datasets, utilizing approximately 1,200, 800, 750, 10,000, and 1,100 instances, respectively.

4.2 Experimental Settings

4.2.1 Evaluation Metrics

To compare our model with previous state-of-the-art (SOTA) models, we evaluate its performance by computing the micro-F1 score across the target domain.

- **INTRA setting**: In traditional NER datasets such as WNUT'17, GUM, I2B2, OntoNotes, and CoNLL'03, distinct tag-set distributions are present. To address this, we generate

[3] https://huggingface.co/datasets/wnut_17
[4] https://gucorpling.org/gum/
[5] https://www.i2b2.org/NLP/DataSets/
[6] https://www.ldc.upenn.edu/
[7] https://huggingface.co/datasets/conll2003

Table 1. Overall Performance Comparison.

	Model	WNUT'17	GUM	I2B2	OntoNotes	CoNLL'03	Avg.
INTRA	ProtoBERT	0.2655	0.1374	0.3433	0.3818	0.3218	0.2900
	NNShot	0.2305	0.0683	0.3844	0.3454	0.3382	0.2734
	ProML	0.2262	0.2336	0.5654	0.2548	0.3424	0.3249
	CONTaiNER	0.2108	0.1328	0.3807	0.2275	0.3199	0.2543
	CLLMFS	**0.5250**	**0.3840**	**0.5800**	**0.5765**	**0.5750**	**0.5281**
	%Improv.	97.74%	64.38%	2.58%	50.99%	67.93%	62.54%
INTER	ProtoBERT	0.2312	0.0920	0.2713	-	0.2917	0.2216
	NNShot	0.2353	0.0634	0.2823	-	0.3280	0.2048
	ProML	0.2456	0.0703	0.2650	-	0.2960	0.2192
	CONTaiNER	0.2291	0.0687	0.3057	-	0.2681	0.2179
	CLLMFS	**0.4579**	**0.2392**	**0.4413**	-	**0.5128**	**0.4128**
	%Improv.	86.44%	160.0%	44.36%	-	56.34%	86.28%

Table 2. Ablation Analysis.

Modules	F1 Score
LLAMA 2 + LoRA	0.375
LLAMA 2 + LoRA + CL	0.377
LLAMA 2 + LoRA + CL + Noise	0.384

Table 3. Impact of LoRA Module Parameter Combinations on F1-score.

W_q	W_k	W_v	W_o	W_{in}	W_{out}	W_{wte}	F1-score
✓							0.281
	✓						0.283
		✓					0.350
✓	✓						0.329
✓		✓					0.370
✓	✓	✓					0.367
✓	✓	✓	✓				0.360
✓	✓	✓	✓	✓			0.368
✓	✓	✓		✓			**0.375**
✓	✓	✓		✓	✓		0.372
✓	✓	✓		✓	✓	✓	0.373
✓	✓	✓	✓	✓	✓	✓	0.352

multiple support sets by sampling from the original training set to train our model within the source domain. These support sets are subsequently employed for predictions on the original test set.

- **INTER (Cross Domain) setting**: In the cross-domain setting, our model is trained on the OntoNotes dataset, serving as the source domain, and subsequently tested on other datasets, constituting the target domain. The tag sets in different datasets are primarily determined by the dataset creators and often do not overlap. For instance, GUM is focused on social media terminology, whereas I2B2 is centered around medical terminology, leading to almost no overlap. In some rare cases where tag overlap occurs, the tags may still represent slightly different concepts (e.g., one dataset might use "place" to denote a neighborhood, while another might use "position" to refer to a city location). In this setting, the training and test set from OntoNotes is split into *N-way K-shot* for training, and the test set consists of the original test sets from various domains, without utilizing their respective training sets.

4.2.2 *Baselines*

To evaluate CLLMFS's effectiveness, we compare it with several state-of-the-art Few-Shot NER models across various datasets and settings:

- **ProtoBERT** [37] simplifies Few-Shot NER using a span-based prototypical network with a pre-computable BERT encoder. It employs token embeddings to create entity prototypes and utilizes l-2 distance for efficient entity prediction during inference.
- **NNShot** [33] adopts a novel token-level nearest neighbor classification approach, distinguishing itself from prototype-based methods by utilizing the proximity of similar samples in an embedding space.
- **ProML** [6] introduces multiple prompt schemas to enrich label semantics and a novel architecture that synergistically integrates these prompts, advancing metric learning in Few-Shot NER.
- **CONTaiNER** [7] utilizes contrastive learning with Gaussian-distributed token embeddings to enhance Few-Shot NER. It focuses on optimizing generalized objectives to improve entity distinction without overfitting to specific domain attributes.

To ensure fair comparisons, we used the optimal parameters from each model's respective code repositories. All models were trained and evaluated on the same datasets, with metrics averaged over five statistical runs for consistency.

4.3 *Performance Comparison*

The performance comparison in Table 1 illustrates CLLMFS's superior effectiveness, achieving new state-of-the-art (SOTA) results. Across different datasets, CLLMFS shows substantial improvements over previous SOTA models, with average relative gains of 62.54% and 86.28% in micro F1 under the **INTRA** and **INTER** settings, respectively.

CLLMFS excels in various challenging scenarios, spanning both within-domain (INTRA) and cross-domain (INTER) NER tasks. Conventional baseline models, such as ProML, face difficulties in adapting to unseen text domains like GUM due to limited prompt design and methodological constraints. CONTaiNER, although effective in few-shot NER, struggles with substantial domain differences between source and target domains.

Despite these challenges, CLLMFS consistently outperforms SOTA models in both within-domain and cross-domain NER tasks, demonstrating robustness and adaptability. Moreover, CLLMFS effectively handles noisy data, as demonstrated in the WNUT'17 dataset, showcasing its suitability for real-world applications with varying data quality.

4.4 Ablation Analysis

We conducted ablation experiments to systematically investigate the impact of each constituent module on the performance of CLLMFS, which comprises three essential modules: Low-Rank Adaptation (LoRA), Contrastive Learning (CL), and Uniform Gaussian Random Noise (Noise). Due to the complexity of computations involved in the LLAMA 2 model without utilizing LoRA technology, the computational resources available were insufficient to execute the model. Consequently, this experiment is excluded from consideration.

As depicted in Table 2, we observed a clear trend of performance improvement with the inclusion of each additional module, which demonstrates the beneficial impact of incorporating Noise in conjunction with LoRA and CL, further bolstering the model's overall performance.

4.5 Influence of LoRA module selections

To enhance entity extraction tasks in low-resource settings, we systematically investigated the impact of LoRA module selections within the CLLMFS architecture. This architecture includes four weight matrices in the self-attention module (W_q, W_k, W_v, W_o), two in the MLP module (W_{in}, W_{out}), and one for word token embeddings (W_{wte}). We applied LoRA to each weight matrix and explored the optimal configurations to maximize performance.

Using 5-fold cross-validation, we ensured the robustness of our findings, averaging performance over five iterations. The results, summarized in Table 3, show that configurations involving LoRA on W_q, W_k, W_v, and W_{in} consistently outperform others. However, adding LoRA to W_o, W_{out} and W_{wte}does not consistently improve performance, as indicated by varying F1-scores across different configurations.

Overall, these findings underscore the importance of careful parameter tuning in optimizing the effectiveness of the LoRA module for few-shot NER tasks. The observed performance variations highlight the intricate interplay between different module parameters and their collective impact on model performance.

5 Discussion

Our proposed CLLMFS framework achieves promising performance in NER task. Different from previous few-shot NER methods, CLLMFS fine-tunes the model and leverages LLM's capabilities by constructing entity SFT data from limited data, enhancing generalization for few-shot NER tasks. Different from previous meta learning methods, CLLMFS leverages abundant semantic information in LLMs, achieving consistent performance across target domains, even with limited source domain samples. Different from previous in-context learning methods, CLLMFS integrates SFT data for fine-tuning and introduces contrastive learning for FS-NER, enhancing boundary awareness and entity recognition accuracy. Please refer to the Appendix B for some study cases.

6 Conclusion

In this paper, we first propose CLLMFS by enhancing the large language model with contrastive learning for few-shot NER. Our method leverages the inherent knowledge within LLM and utilizes LoRA for supervised fine-tuning. By integrating contrastive learning, CLLMFS enhances LLM's ability of boundary awareness and entity extraction accuracy. Our approach has achieved state-of-the-art performance on multiple datasets with limited labeled data. The cross-domain experiment results confirm that the strong transfer capabilities of CLLMFS across different domains. In the future, we will concentrate on named entity recognition and extend our current work to relation extraction.

References

[1] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, and S. Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298.

[2] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

[3] M. M. Bejani and M. Ghatee. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, pages 1–48, 2021.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] Y. Chen, Y. Zheng, and Z. Yang. Prompt-based metric learning for few-shot NER. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7199–7212, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.451. URL https://aclanthology.org/2023.findings-acl.451.

[7] S. S. S. Das, A. Katiyar, R. J. Passonneau, and R. Zhang. Container: Few-shot named entity recognition via contrastive learning. *arXiv preprint arXiv:2109.07589*, 2021.

[8] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In L. Derczynski, W. Xu, A. Ritter, and T. Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4418. URL https://aclanthology.org/W17-4418.

[9] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL http://arxiv.org/abs/2002.06305. cite arxiv:2002.06305.

[10] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[11] J. Fang, X. Wang, Z. Meng, P. Xie, F. Huang, and Y. Jiang. Manner: A variational memory-augmented model for cross domain few-shot named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4261–4276, 2023.

[12] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[14] J. Huang, C. Li, K. Subudhi, D. Jose, S. Balakrishnan, W. Chen, B. Peng, J. Gao, and J. Han. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 10408–10423, 2021.

[15] Q. Lv, M. Ding, Q. Liu, Y. Chen, W. Feng, S. He, C. Zhou, J. Jiang, Y. Dong, and J. Tang. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1150–1160, 2021.

[16] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[18] N. M. Shazeer. Glu variants improve transformer. *ArXiv*, abs/2002.05202, 2020. URL https://api.semanticscholar.org/CorpusID:211096588.

[19] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:309759.

[20] A. Stubbs and Ö. Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

[21] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of biomedical informatics*, 58:S67–S77, 2015.

[22] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864, 2021. URL https://api.semanticscholar.org/CorpusID:233307138.

[23] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[26] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[27] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *Neural Information Processing Systems*, 2016. URL https://api.semanticscholar.org/CorpusID:8909022.

[28] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, and Z. Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 968–977, 2019.

[29] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo. Knowledge graph convolutional networks for recommender systems. In *The world wide web conference*, pages 3307–3313, 2019.

[30] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang. Gpt-ner: Named entity recognition via large language models, 2023.

[31] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958, 2019.

[32] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, et al. Ontonotes release 5.0 ldc2013t19. linguistic data consortium, philadelphia, pa (2013), 2013.

[33] Y. Yang and A. Katiyar. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*, 2020.

[34] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, and Q. V. H. Nguyen. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1294–1303, 2022.

[35] A. Zeldes. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.

[36] B. Zhang and R. Sennrich. Root mean square layer normalization. *ArXiv*, abs/1910.07467, 2019. URL https://api.semanticscholar.org/CorpusID:113405151.

[37] Y. Zhang and H. Fang. Less is more: A prototypical framework for efficient few-shot named entity recognition. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–46. Springer, 2023.

Appendix A Different entity types in SFT

A.1 Location

```
{ "instruction": "Please extract the entity of
location in the input sentence given below , the
entity of place refers to the entity that represents
the name or place of a specific location in the
input sentence.", "input": "The hijackers told the
crew they had grenades and other explosives and
threatened to blow up the plane if they were not
taken to London .", "output": "<im_start> I can
extract entities for you, the extracted entities are
<<< London >>> <im_end>" }
```

A.2 Organization

```
{ "instruction": "Please extract the entity of
organization in the input sentence given below ,
the entity of organization refers to the entity
that represents a specific organization in the
input sentence.", "input": "There 's no problem
whatsoever , \" he told Reuters . He said the
lifestyle associated with being Miss Universe
could make routine exercise difficult .", "output":
"<im_start> I can extract entities for you, the
extracted entities are <<< Reuters >>> <im_end>" }
```

A.3 Law

```
{ "instruction": "Please extract the entity of Law
in the input sentence given below, the entity of
Law refers to the entity that represents a rule or
system of rules recognized by a country or community
in the input sentence.", "input": "but it seems
to me unlikely on somebody that would be more of
a strict constructionist on the constitution /.",
"output": "<im_start> I can extract entities for
you, the extracted entities are <<< constitution >>>
<im_end>" }
```

A.4 Work of Art

```
{ "instruction": "Please extract the entity of Work
of Art in the input sentence given below, the entity
of Work of Art refers to the entity that represents
a creative or artistic product or performance in
the input sentence.", "input": "They were extremely
disappointed with the sound on Autumn Moon , so they
decided to go to Australia to redo it .", "output":
"<im_start> I can extract entities for you, the
extracted entities are <<< Autumn Moon >>> <im_end>"
}
```

A.5 Product

```
{ "instruction": "Please extract the entity
of product in the input sentence given below ,
the entity of product refers to the entity that
represents the name or category of a specific
product in the input sentence.", "input": " New
#VideoRecorder #7 : Flip MinoHD Video Camera -- 8
GB , 2 Hour Record Time : Flip MinoHD Video Camera
-- ... http://dlvr.it/5P5kV #cheaptweet", "output":
"<im_start> I can extract entities for you, the
extracted entities are <<< Flip MinoHD >>> <<< Flip
MinoHD >>> <im_end>" }
```

A.6 Corporation

```
{ "instruction": "Please extract the entity of
corporation in the input sentence given below ,
the entity of corporation refers to the entity that
represents the name or organization of a specific
corporation in the input sentence.", "input": "
Jackie joins Twitter 50 years after JFK campaign [
AP : US News ] http://bzbx.us/uEf #50 #Campaign via
buzzbox . com", "output": "<im_start> I can extract
entities for you, the extracted entities are <<<
Twitter >>> <im_end>" }
```

Appendix B Study Cases

B.1 Few-shot cases

The term "None" in Table A1 indicates that the model produced no output. The results demonstrate that our approach exhibits superior semantic understanding in low-resource NER tasks, effectively identifying entities such as 'Andrew' and 'Little'. CLLMFS demonstrates the ability to extract the complete person name entity from the sentence, while other methods are best at extracting features 'Andrew' . Consequently, we assert that CLLMFS offers superior semantic understanding and is less susceptible to being influenced solely by the inherent meaning of the word 'Little'. This advantage can be attributed to the incorporation of LLAMA2 in our methodology.

In the case of Table A2, it can be observed that NNshot and ProtoBERT methods exhibit limited performance in recognizing specialized noun entities, indicating their constrained generalization capability and susceptibility to neighboring characters' influence. CONTaiNER and ProML successfully identify the target entities but fail to capture the semantics of specialized nouns, resulting in the extraction of irrelevant entities. In contrast, our method, based on a large-model architecture, accurately identifies specialized noun entities like 'Trump administration' without extracting any erroneous entities.

B.2 Contrastive learning cases

In Table A3, Lack of contrastive learning in the model results in poor boundary perception, thereby leading to inadequate semantic comprehension and increased likelihood of over-extraction or under-extraction in low-resource entity extraction tasks.

In the case of Table A4, it is evident that without contrastive learning, not only was the entity 'the world' extracted, but also other entities such as 'some great musicians'. By introducing contrastive learning as an auxiliary task, CLLMFS have effectively improved the internal representations of the large model, resulting in more precise entity extraction.

B.3 Transfer learning cases

The examples below are all trained on the OntoNotes source data and evaluated on the I2B2, WNUT'17, GUM, and CoNLL'03 datasets, respectively. Across these datasets, our approach demonstrates robust cross-domain NER capabilities.

In the medical domain dataset I2B2 (Table A5), our method excels in identifying relevant entities, while CONTaiNER and ProML, although capable of extraction, introduce too many erroneous entities. In the social media dataset WNUT'17 (Table A6), our method effectively identifies location entities, whereas ProML successfully recognizes them but also extracts other entity information. In the GUM dataset (Table A7), our method uniquely identifies relevant entities correctly. In the news dataset CoNLL'03 (Table A8), only our method and NNShot can accurately identify the journalistic terminology 'reuters'. In summary, this demonstrates the strong cross-domain entity extraction capability of CLLMFS.

Table A1. Case A1.

Input	True, but I imagine it would be a lot lower and as I pointed out to Andrew Little would be cheaper than [eliminating fees.
label	’Andrew’, ’Little’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’Andrew’	’but’, ’it’, ’be’, ’as’, ’out’	’Andrew’	None	’Andrew’, ’Little’

Table A2. Case A2.

Input	/ . Trump administration rolls back protections for people in default on student loans https://t.co/YPe2DRsjLe
label	’Trump’, ’administration’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’/’, ’default’	’.’, ’Trump’, ’back’, ’protections’, ’for’, ’default’, ’on’, ’loans’	’Trump’, ’protections’, ’default’, ’loans’	’administration’, ’loans’	’Trump’, ’administration’

Table A3. Case A3.

Input	topless sunbathing is common by western women at many tourist beaches .
label	’many tourist beaches’

Model	CLLMFS without CL	CLLMFS
Output	’tourist beaches’	’many tourist beaches’

Table A4. Case A4.

Input	i worked with some great musicians and performed around the world .
label	’the world’

Model	CLLMFS without CL	CLLMFS
Output	’the world’, ’some great musicians’	’the world’

Table A5. Case A5.

Input	she came from turlock 10 years ago to pecos to be with her daughter who is a surgeon in the area
label	’surgeon’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’she’, ’came’, ’ago’, ’be’, ’is’	None	’pecos’, ’her’, ’daughter’, ’who’, ’is’, ’a’, ’surgeon’, ’in’, ’area’	’pecos’, ’with’, ’her’, ’daughter’, ’who’, ’a’, ’surgeon’, ’in’, ’area’	’surgeon’

Table A6. Case A6.

Input	i looked this up yesterday actually ; the average household income for auckland in 2016 was $ 104 k .
label	’auckland’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’I’, ’household’, ’in’, ’2016’	’household’, ’2016’	’this’, ’Auckland’, ’104’, ’k’, ’.’	None	’auckland’

Table A7. Case A7.

Input	most travellers will arrive in York by road (car or bus) or rail from other parts of the uk or an airport .
label	’most travellers’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’in’, ’York’, ’by’, ’car’, ’parts’, ’of’, ’an’	’York’, ’car’, ’other’	’bus’, ’rail’, ’from’, ’other’, ’parts’, ’of’, ’or’	’York’, ’road’, ’rail’, ’from’, ’other’, ’parts’	’most travellers’

Table A8. Case A8.

Input	he told reuters he had needed to speak to her before she left wellington later on friday .
label	’reuters’

Model	ProtoBERT	NNShot	ProML	CONTaiNER	CLLMFS
Output	’reuters’, ’wellington’	’reuters’	None	None	’reuters’