Heterogeneous Collaboration and Instance-Adaptive Constraints for Few-Shot Nested NER

Zhiqiang Liu[1], Dan Wu[1], Juncheng Jia[1,*], Yang Yang[2,3,*]
[1]School of Computer Science & Technology, Soochow University, Suzhou, China
[2]Computing Science and Artificial Intelligence College, Suzhou City University
[3]Suzhou Key Lab of Multi-modal Data Fusion and Intelligent Healthcare, Suzhou, China
{zqliu, dwudwu}@stu.suda.edu.cn, jiajuncheng@suda.edu.cn, yyang@szcu.edu.cn

***Abstract*—Few-Shot Nested Named Entity Recognition (NER) is dedicated to parsing complex hierarchical entity structures in extremely low-resource scenarios. However, the two existing mainstream approaches face distinct challenges: discriminative models are highly prone to overfitting in few-shot settings, while Generative Large Language Models (LLMs), constrained by their vast decoding spaces, frequently suffer from issues of entity boundary drift and hallucination. To this end, this paper proposes a heterogeneous collaborative framework, HeCoNER, designed to leverage discriminative models to extract entity structural features to compress the decoding space of LLMs. Specifically, we first guide the discriminative model to robustly extract entity boundary features via Dynamic Prompts. Next, we utilize a Heterogeneous Feature Alignment module to filter out noise from the entity boundary features and map them into the semantic space of the LLM. Serving as instance-adaptive constraints, these features are finally injected into the LLM, transforming open-ended generation into a constrained decoding process. Experimental results demonstrate that HeCoNER achieves significant performance improvements across multiple benchmark datasets, validating the effectiveness of the framework in complex nested scenarios.**

***Index Terms*—Few-Shot Learning, Named Entity Recognition, Large Language Model, Attention Mechanism**

I. Introduction

Named Entity Recognition (NER) aims to extract entities with specific meanings from unstructured text and serves as a core component in building knowledge graphs and question answering systems [1]. While traditional methods focus on flat entities, nested structures prevail in domains like biomedicine and finance [2]. For instance, in "[[Peking University] Affiliated Hospital]", both the inner "Peking University" and the outer full institution name need to be recognized simultaneously. This nesting challenges standard sequence labeling models [3]. Moreover, given the scarcity and high cost of domain-specific data [4], Few-Shot Nested NER has emerged as a research hotspot.

Although existing research has achieved progress in the field of few-shot nested NER, mainstream discriminative and generative methods still face severe theoretical bottlenecks. For discriminative models [3], [5]–[7] represented by BERT [8], the approach relies on full-parameter fine-tuning on the target domain data. However, in extremely low-resource scenarios (e.g., 5-shot), the optimization of such a high-dimensional parameter space is highly prone to overfitting: the model tends to memorize superficial features rather than learning underlying patterns [9], leading to catastrophic forgetting of pre-trained general knowledge [10], thereby severely weakening its generalization ability on complex nested structures.

Secondly, the Masked Language Modeling of BERT models tends to drive the model to capture semantic co-occurrence relationships between words, causing semantically similar words to cluster tightly in the feature space [11]. However, nested NER relies more heavily on syntactic structures and boundary features (such as prepositions, punctuation, and case changes) to determine the start and end positions of entities, rather than relying solely on the meanings of the words themselves. The original attention mechanism of BERT often overlooks these structural clues crucial for boundary localization, thereby leading to a significant objective misalignment between the "semantic bias" acquired by the model and the "structural perception" required for nested NER [12].

On the other hand, generative Large Language Models (LLMs) (e.g., Llama [13]), while excelling in semantic understanding and zero-shot reasoning, are constrained by their probabilistic generation paradigm and often exhibit a characteristic of "strong semantics, weak structure". Due to their vast decoding space, in few-shot scenarios, these models struggle to precisely locate nested entity boundaries [14], frequently suffering from boundary drift (e.g., including redundant prepositions or truncating entities) or hallucination (e.g., generating entity words not present in the sentence), leading to a significant decline in recognition accuracy [15].

In view of the aforementioned issues, we advocate combining the advantages of discriminative models in feature encoding with the advantages of generative models in semantic reasoning. However, due to significant discrepancies between the feature spaces of the two, simply using linear layers or MLPs for feature alignment is difficult to train sufficiently under few-shot conditions, thereby introducing a large amount of noise into the generative model.

To this end, this paper proposes a Heterogeneous Collaborative Framework (HeCoNER) tailored for few-shot nested NER.

*Corresponding authors: Juncheng Jia and Yang Yang

This work was supported by the Fundamental Research Funds for Suzhou Key Lab of Multi-modal Data Fusion and Intelligent Healthcare (25SZZD02) and the Suzhou Sports Bureau Sports Science Research Project (TY2025-203).

The core idea of this framework is to utilize the discriminative model to guide the decoding process of the generative model while preserving pre-trained knowledge. Specifically, we first adopt a parameter freezing strategy and introduce learnable dynamic prompt vectors to guide the BERT encoder to pay more attention to syntactic structures and boundary words rather than solely focusing on semantics, thereby extracting robust structural features without fine-tuning the BERT model parameters. Subsequently, to address the feature space inconsistency between BERT and the LLM and the noise issues caused by feature alignment under few-shot conditions, we design an attention-based heterogeneous feature alignment module to filter and map the extracted discriminative features into the semantic space of the generative model. Finally, we explicitly inject these features containing rich structural information into the input of the LLM as instance-adaptive constraints. This approach aims to provide the LLM with an instance-specific contextual reference, subjecting it to explicit boundary guidance during the decoding process, transforming open-ended generation into a controlled structural restoration process and ensuring precise recognition of complex nested entities even with few samples. Code is available at: https://github.com/LZQ12346/HeCoNER. Our contributions include:

•We propose HeCoNER, an end-to-end framework based on heterogeneous model collaboration. By employing dynamic prompts, we effectively circumvent the risk of overfitting caused by few-shot fine-tuning of discriminative models, while utilizing the semantic reasoning capability of LLMs to compensate for deficiencies in parsing complex nested structures.

•We design a heterogeneous feature alignment and instance-adaptive constraint injection mechanism. This mechanism filters noise features from the discriminative model and transforms structural features into instance-specific contexts, reconstructing the open-ended generation of the LLM into a controlled decoding process, thereby significantly improving the precision of boundary localization.

•Significant performance improvements are achieved on multiple benchmark datasets, demonstrating exceptional robustness particularly in deep nested scenarios.

II. Related Work

A. Discriminative Model-based Few-Shot Nested NER

Existing research on Nested NER primarily follows the discriminative paradigm, with mainstream methods including span-based [16], hypergraph-based [17], [18], and layered models [19], [20]. These methods widely adopt pre-trained language models such as BERT as feature encoders. Their core success factor lies in the fact that, in rich-resource scenarios, strong supervision signals from full-scale data can drive the model to undergo deep parameter reconstruction, thereby effectively adapting and transforming the general representations of the pre-trained model—which originally focus on "semantic co-occurrence"—into "structural boundary" features crucial for nested entities. However, this semantic-structural feature adaptation relying on large-scale data is difficult to sustain under few-shot conditions, causing the model to be forced to reduce loss through mechanical memorization of superficial features, thereby inevitably triggering overfitting and catastrophic forgetting [7]. To alleviate the optimization challenges brought by data scarcity, recent research has begun to explore adaptive strategies. Span-based methods [5], [21] such as FIT [3] adopt a multi-stage decomposition strategy to reduce learning difficulty, while meta-learning-based methods [6], [7] are dedicated to transferring general features from the source domain. Nevertheless, these methods inherently fail to break away from the dependency on parameter updates for discriminative models in the few-shot target domain: few-shot training on various sub-tasks in multi-stage strategies leads to Error Cascade Propagation of overfitting errors [4]; while meta-learning methods introduce prior knowledge, they still struggle to completely eliminate the parameter optimization instability caused by data sparsity during the target domain adaptation process, failing to fundamentally circumvent the risk of overfitting.

B. Generative Model-based Few-Shot Nested NER

Generative LLMs, represented by LLaMA [13], leverage their exceptional semantic understanding capabilities to provide a new paradigm for information extraction. In nested NER tasks, mainstream research typically adopts a Sequence Linearization strategy, utilizing T5 [22], or prompt-guided LLMs [13], [23] to transform complex layered entity structures into flattened label sequences or specific structured strings for autoregressive prediction [4]. When resources are sufficient, this approach can utilize a large number of samples to train the model to align entity structural features to specific output formats, thereby effectively circumventing the complex decoder designs found in discriminative models [24]. However, this strong reliance on "format alignment" becomes a fatal weakness in few-shot scenarios. To alleviate the model's unfamiliarity with the output format, the T5-based in-context learning framework EnDe [4] attempts to explicitly "teach" the model to follow a specific nested decoding paradigm by retrieving similar entity demonstrations. Although this retrieval-augmented strategy standardizes the generation format to a certain extent, it inherently fails to alter the open-ended probabilistic prediction characteristic of generative models based on the full vocabulary. The vast search space makes the model highly prone to detaching from the constraints of the original text [14], which not only leads to drift in boundary localization (e.g., absorbing irrelevant words) but also triggers generative hallucinations, where the model erroneously generates entity content not present in the original text amidst an uncertain probability distribution.

III. Method

As shown in Fig. 1, our proposed HeCoNER framework operates through the collaboration of three core components: (1) a Dynamic Prompt-Guided BERT Encoder, designed to extract robust entity structural features and boundary signals; (2) Heterogeneous Feature Alignment & Denoising, responsible for bridging the gap between heterogeneous feature spaces

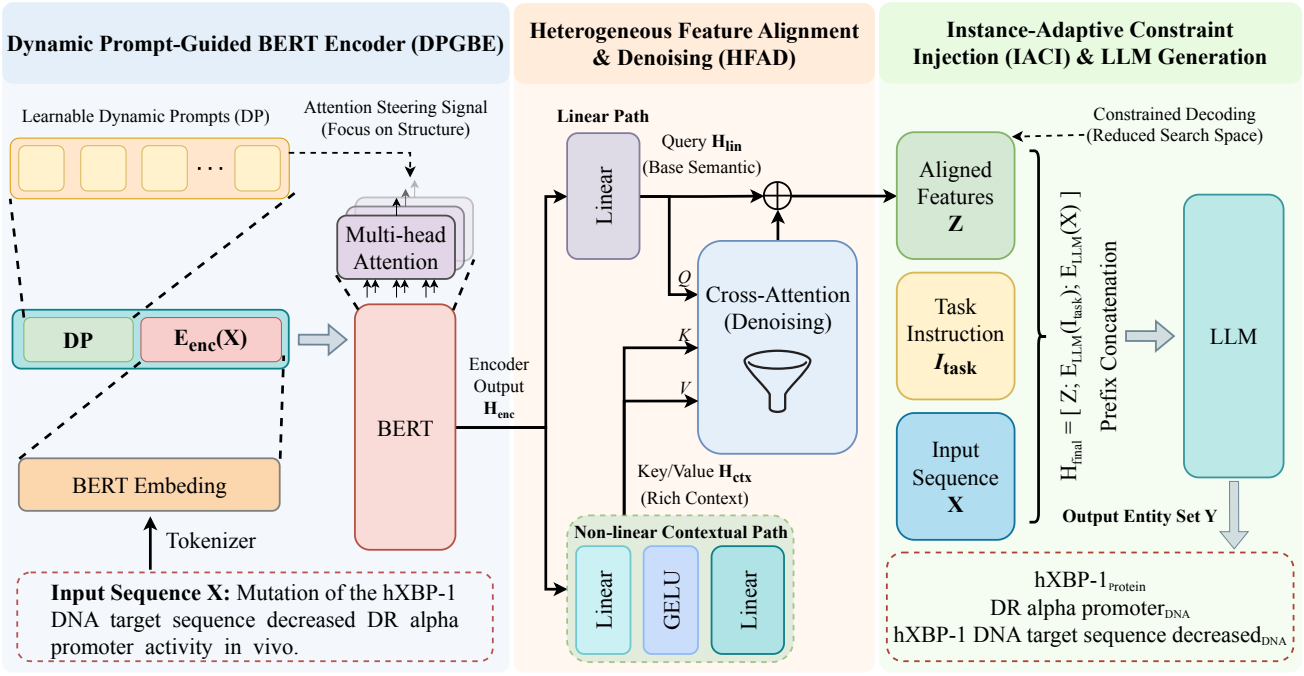


Fig. 1. The overall architecture of our proposed HeCoNER framework.

and simultaneously filtering out redundant noise extracted from upstream; and (3) Instance-Adaptive Constraint Injection, which precisely guides the generation path of entities by compressing the decoding search space of the LLM.

A. Task Definition

Formally, we reformulate the Nested NER task as a text-to-text generation task. Given an input sentence sequence $X = \{x_1, x_2, \ldots, x_N\}$, the goal of NER is to extract an entity set $Y = \{e_k\}_{k=1}^{M}$, where each entity is represented by a tuple $e_k = \langle x_{s_k:d_k}, t_k \rangle$, with s_k, d_k, and t_k corresponding to the start index, end index, and entity type, respectively. Unlike traditional Flat NER, which constrains the spans of any two distinct entities e_i and e_j to be mutually disjoint (i.e., $[s_i, d_i] \cap [s_j, d_j] = \emptyset$), Nested NER relaxes this restriction, allowing the spans of different entities to have overlapping or inclusion relationships, meaning there may exist e_i, e_j such that $[s_i, d_i] \cap [s_j, d_j] \neq \emptyset$. Few-Shot Nested NER aims to utilize an extremely limited support set (i.e., K samples per entity class) to learn complex hierarchical structures from sparse supervision and achieve robust generalization on nested entities in the query set.

B. Dynamic Prompt-Guided BERT Encoder

To alleviate the misalignment between the semantic-dominant nature of pre-trained models and the structural requirements of nested entity recognition, and to circumvent the risk of overfitting induced by few-shot fine-tuning, we propose the Dynamic Prompt-Guided BERT Encoder (DPGBE) strategy. Specifically, based on freezing the BERT encoder parameters θ_{fixed}, we introduce a set of learnable continuous dynamic prompt vectors $\mathbf{DP} \in \mathbb{R}^{m \times d_{enc}}$. Given the input text sequence X, we prefix the dynamic prompt vectors to the text embeddings $\mathbf{E}_{enc}(X)$ to construct the augmented input representation $\mathbf{H}_{in}$:

$$\mathbf{H}_{in} = [\mathbf{DP}\,;\, \mathbf{E}_{enc}(X)] \tag{1}$$

During the encoding process, the dynamic prompt vectors $\mathbf{DP}$ serve as an Attention Steering Signal. Through the interaction of the multi-head self-attention mechanism, the dynamic prompt vectors alter the distribution of attention scores, thereby suppressing the model's excessive focus on pure semantics and activating features within the frozen parameters that are sensitive to lexical boundaries:

$$\mathbf{H}_{enc} = \text{Softmax}\left(\frac{(\mathbf{H}_{in}\mathbf{W}^{Q})(\mathbf{H}_{in}\mathbf{W}^{K})^{T}}{\sqrt{d}}\right)(\mathbf{H}_{in}\mathbf{W}^{V}) \tag{2}$$

This mechanism extracts structural features adaptive to the nested NER task through the guidance of dynamic prompt, without compromising the pre-trained general representations.

C. Heterogeneous Feature Alignment & Denoising

Given the significant differences in distribution and dimension between the discriminative feature space of BERT and the generative feature space of the LLM, simple linear or MLP mappings under few-shot conditions are highly prone to propagating noise from the BERT encoder side. To this end, we design a Heterogeneous Feature Alignment & Denoising (HFAD) module, which aims to map the low-dimensional features $\mathbf{H}_{\text{enc}} \in \mathbb{R}^{L \times d_{enc}}$ output by BERT to the high-dimensional space d_{llm} of the LLM and perform dynamic

SYSTEM: The task is to label named entities from in the given sentence and the entity should be chosen in [*OPTIONS*]

OPTIONS: ["ORG", "GPE", "PER", "LOC", "FAC", "VEH", "WEA"]

INSTRUCTION: Please answer in the format ["entity": "name", "entity_type": "type"], Here is the sentence:

Fig. 2. An illustrative example of the text-based task instruction I_{task} constructed for the ACE2004 dataset.

filtering to preserve effective entity boundary features. This module adopts a dual-path projection architecture:

First, we construct a Non-linear Contextual Path. This path utilizes a two-layer perceptron (MLP) with the GELU activation to project features to an intermediate dimension d_{mid} and then map them to the target dimension d_{llm}. This process aims to enrich the semantic representation of features through non-linear transformations, and its output $\mathbf{H}_{\text{ctx}}$ serves as the Key and Value in the subsequent attention mechanism:

$$\mathbf{H}_{\text{ctx}} = \text{GELU}(\mathbf{H}_{\text{enc}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_e \times d_{mid}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{mid} \times d_{llm}}$ are learnable projection matrices.

Second, we construct a Linear Projection Path. This path performs a linear transformation on $\mathbf{H}_{\text{enc}}$ to generate baseline features $\mathbf{H}_{\text{lin}}$ that preserve the original distributional characteristics, serving as the Query in the attention mechanism:

$$\mathbf{H}_{\text{lin}} = \mathbf{H}_{\text{enc}}\mathbf{W}_p + \mathbf{b}_p \tag{4}$$

To achieve dynamic denoising, we feed the outputs of both paths into a Cross-Attention mechanism. This mechanism utilizes the linear baseline features $\mathbf{H}_{\text{lin}}$ to dynamically probe for effective structural features within the non-linear representations $\mathbf{H}_{\text{ctx}}$, filtering noise via cross-attention. Finally, we apply a residual connection between the attention output and the linear features to obtain the aligned high-dimensional features $\mathbf{Z}$:

$$\mathbf{Z} = \text{Softmax}\left(\frac{\mathbf{H}_{\text{lin}}(\mathbf{H}_{\text{ctx}})^T}{\sqrt{d_{llm}}}\right)\mathbf{H}_{\text{ctx}} + \mathbf{H}_{\text{lin}} \tag{5}$$

Through this residual fusion strategy, we ensure that the representations input into the LLM possess a high signal-to-noise ratio and contain entity structural information.

D. Instance-Adaptive Constraint Injection

To alleviate the hallucination and boundary drift issues caused by the lack of structural priors in generative models under few-shot conditions, and to overcome the limitations of traditional static prompts in capturing dynamic syntactic structures, we propose the Instance-Adaptive Constraint Injection (IACI) strategy. This module adopts a "Dual-Prompting Strategy", synergistically injecting natural language instructions and structural features into the LLM. Specifically, we first define a textual task instruction I_{task} (e.g., Fig. 2) to clarify the generation objective. Subsequently, we concatenate the aligned feature matrix $\mathbf{Z}$ output from the previous stage with the task instruction embeddings $\mathbf{E}_{\text{LLM}}(I_{\text{task}})$ and the input text embeddings $\mathbf{E}_{\text{LLM}}(X)$ along the sequence dimension to construct the final guided input representation $\mathbf{H}_{\text{final}}$:

$$\mathbf{H}_{\text{final}} = [\mathbf{Z}\,;\,\mathbf{E}_{\text{LLM}}(I_{\text{task}})\,;\,\mathbf{E}_{\text{LLM}}(X)] \tag{6}$$

In this architecture, the text instruction I_{task} provides explicit task guidance, while the feature matrix $\mathbf{Z}$ serves as a critical instance-specific Contextual Reference. $\mathbf{Z}$ supplements fine-grained boundary information in the latent space that discrete instructions cannot describe, explicitly compressing the model's generation search space. Based on this, the generation probability distribution of the LLM is reformulated as a Constrained Decoding process:

$$P(Y|X) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, \mathbf{H}_{\text{final}}) \tag{7}$$

Through this injection mechanism, the model effectively combines explicit task definitions with implicit structural constraints, precisely localizing the complex boundaries of nested entities under few-shot conditions and achieving robust inference output.

IV. Experiments

A. Dataset and Experimental Setup

We evaluate on ACE2004 [26], ACE2005 [27], and GENIA [28] under 5/10/20-shot settings. Adopting the recognized sampling method from previous research [29], we conduct 5 independent sampling and training runs, reporting the average F1 score across these experiments. Experiments are implemented using PyTorch 2.5.1 and Transformers 4.50 on a single RTX 3090 (24GB) GPU. Regarding model configuration, we freeze the BERT [8] parameters, fine-tune the linear layers of the LLM via LoRA [30] (rank $r = 32$, $\alpha = 64$, Dropout=0.1), and perform full-parameter fine-tuning on the HAFD module; additionally, the DP vector length is fixed at 10. Training employs the AdamW optimizer (learning rate $5e-5$, batch size 1) for a total of 30 epochs, with the optimal validation checkpoint selected for testing. To ensure fair comparison, we group experiments based on parameter scale. In the lightweight ($<$ 1B) setting, we select Qwen3-0.6B [23] (as only the Qwen series contains tiny models). In the regular ($>$ 8B) setting, we employ Llama3-8B [13] due to its superior performance in English, which aligns with our English datasets. Evaluation follows the Exact Match standard (requiring strict boundary and type alignment).

B. Baselines

To comprehensively evaluate model performance, we select recent SOTA models in nested NER tasks as baselines and categorize them into two groups based on parameter scale. The first group includes sub-1B parameter methods based on BERT or T5: SDNet [25]; ESD [21], which employs span-level matching with prototype aggregation; FIT [3], which combines soft prompts and contrastive learning for span classification;

TABLE I

PERFORMANCE COMPARISON OF VARIOUS METHODS ACROSS THREE NESTED NER DATASETS. RESULTS ARE REPORTED IN F1 SCORES (MEAN±STD %). THE SYMBOL † DENOTES BERT-BASED BASELINES, AND **BOLD** HIGHLIGHTS THE BEST RESULTS.

Model	ACE2004			ACE2005			GENIA		
	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot
Parameter Size $<$ 1B									
SDNet† [25]	$20.55_{\pm 4.46}$	$34.82_{\pm 4.71}$	$42.87_{\pm 2.13}$	$22.03_{\pm 6.12}$	$32.20_{\pm 4.89}$	$43.00_{\pm 3.55}$	$17.46_{\pm 6.97}$	$19.03_{\pm 7.07}$	$33.27_{\pm 3.71}$
ESD† [21]	$19.25_{\pm 5.74}$	$42.75_{\pm 5.11}$	$52.17_{\pm 3.76}$	$31.57_{\pm 6.45}$	$38.81_{\pm 7.04}$	$50.30_{\pm 3.37}$	$25.03_{\pm 9.88}$	$35.23_{\pm 4.96}$	$47.22_{\pm 4.36}$
FIT† [3]	$35.87_{\pm 4.92}$	$44.88_{\pm 4.82}$	$53.92_{\pm 2.99}$	$37.74_{\pm 5.33}$	$42.25_{\pm 10.65}$	$52.71_{\pm 2.55}$	$34.43_{\pm 9.06}$	$44.95_{\pm 3.38}$	$51.26_{\pm 3.96}$
EnDe [4]	40.15	48.43	55.45	44.42	46.10	54.98	39.85	48.27	54.39
TECA† [5]	$40.18_{\pm 4.32}$	$50.50_{\pm 1.81}$	$58.19_{\pm 1.57}$	$39.19_{\pm 6.24}$	$47.36_{\pm 2.87}$	$55.99_{\pm 3.50}$	$34.60_{\pm 5.88}$	$45.65_{\pm 3.71}$	$52.69_{\pm 3.92}$
HeCoNER (Qwen3-0.6B)	$\mathbf{45.85}_{\pm 3.34}$	$\mathbf{53.54}_{\pm 5.68}$	$\mathbf{58.42}_{\pm 2.72}$	$\mathbf{46.51}_{\pm 4.13}$	$\mathbf{52.86}_{\pm 3.19}$	$\mathbf{56.90}_{\pm 2.16}$	$\mathbf{47.23}_{\pm 3.73}$	$\mathbf{53.45}_{\pm 3.21}$	$\mathbf{55.16}_{\pm 2.42}$
Parameter Size $>$ 8B									
Llama3-8B [13]	$51.35_{\pm 3.88}$	$52.68_{\pm 6.44}$	$59.09_{\pm 3.07}$	$47.51_{\pm 5.10}$	$50.86_{\pm 6.71}$	$58.90_{\pm 3.25}$	$55.88_{\pm 3.38}$	$57.07_{\pm 3.51}$	$58.67_{\pm 2.42}$
Qwen3-8B [23]	$49.43_{\pm 3.91}$	$52.54_{\pm 4.72}$	$57.77_{\pm 3.82}$	$45.89_{\pm 4.41}$	$49.79_{\pm 3.65}$	$55.41_{\pm 2.88}$	$51.94_{\pm 3.35}$	$52.76_{\pm 3.23}$	$57.32_{\pm 2.69}$
T5-xxl [22]	$50.26_{\pm 4.19}$	$51.93_{\pm 6.38}$	$56.69_{\pm 4.61}$	$45.98_{\pm 4.32}$	$49.68_{\pm 3.97}$	$56.32_{\pm 3.88}$	$53.19_{\pm 4.53}$	$55.63_{\pm 3.94}$	$58.86_{\pm 3.26}$
HeCoNER (Llama3-8B)	$\mathbf{52.19}_{\pm 2.97}$	$\mathbf{55.54}_{\pm 4.12}$	$\mathbf{60.20}_{\pm 2.54}$	$\mathbf{47.99}_{\pm 3.85}$	$\mathbf{54.88}_{\pm 3.04}$	$\mathbf{61.13}_{\pm 1.99}$	$\mathbf{56.84}_{\pm 3.01}$	$\mathbf{60.23}_{\pm 2.90}$	$\mathbf{61.24}_{\pm 2.16}$

TABLE II

ABLATION STUDY RESULTS ON THE GENIA DATASET (F_1). NOTE THAT "W/O IACI" DENOTES THE REMOVAL OF ALL PROPOSED MODULES, REPRESENTING A BASELINE WHERE THE QWEN3-0.6B BACKBONE IS FINE-TUNED SOLELY VIA LORA.

Methods	5-shot	10-shot	20-shot
HeCoNER (Qwen3-0.6B)	**47.23**	**53.45**	**55.16**
- w/o IACI	45.26 ↓1.97	48.90 ↓4.55	53.41 ↓1.75
- w/o Dynamic Prompt	46.85 ↓0.38	52.96 ↓0.49	55.03 ↓0.13
- w/o HFAD:			
Replaced by Linear	40.15 ↓7.08	43.58 ↓9.87	48.01 ↓7.15
Replaced by MLP	43.83 ↓3.40	47.43 ↓6.02	51.48 ↓3.68

TABLE III

COMPARISON OF BERT TRAINING STRATEGIES UNDER THE 1-SHOT SETTING ($\text{MEAN}_{\pm\text{STD}}$). ALL RESULTS ARE DERIVED FROM THE HECONER FRAMEWORK WITH A FIXED QWEN3-0.6B BACKBONE, VARYING STRICTLY THE TUNING METHOD OF THE BERT ENCODER.

BERT Training Strategy	ACE2004	ACE2005	GENIA
Full Fine-tuning	$39.89_{\pm 10.04}$	$40.02_{\pm 9.38}$	$41.24_{\pm 11.25}$
LoRA Fine-tuning	$40.68_{\pm 7.26}$	$41.32_{\pm 6.91}$	$42.39_{\pm 8.05}$
Frozen (Ours)	$\mathbf{42.36}_{\pm 4.86}$	$\mathbf{43.18}_{\pm 4.17}$	$\mathbf{44.71}_{\pm 4.97}$

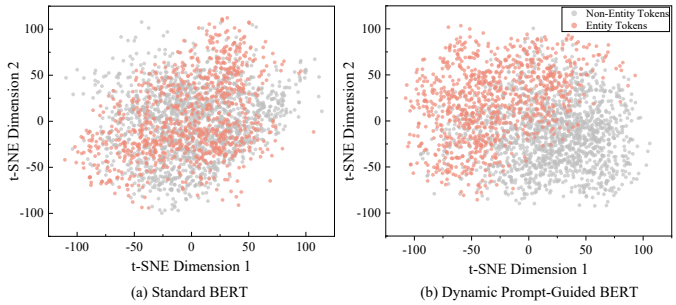
TECA [5], featuring enhanced contextual attention for nested entities; and EnDe [4], a T5-based model utilizing retrieval and contrastive learning for demonstration selection. The second group comprises generative LLMs ($>$ 8B params): Llama3-8B [13], Qwen3-8B [23], and T5-xxl [22]. To ensure fair comparison under limited resources, we uniformly fine-tune these models using the same LoRA strategy as our method (see Sec. IV-A).

C. Main Results

As shown in Table I, within comparable parameter scales, the HeCoNER framework achieves performance significantly superior to comparative baselines across all few-shot settings, while exhibiting exceptional stability as evidenced by the standard deviation across multiple random experiments. (1) Superiority over SLMs ($<$1B): Particularly against BERT-based baselines such as SDNet, ESD, and FIT, our method demonstrates results characterized by low variance and high performance. This intuitively validates our design rationale regarding mitigating the overfitting risks associated with few-shot fine-tuning of BERT: our Dynamic Prompt mechanism successfully serves as an attention steering signal, guiding the BERT encoder to focus more on syntactic structures and boundary words, thereby achieving stable structural knowledge extraction without compromising pre-trained knowledge. (2) Advantage over LLMs ($>$8B): Simultaneously, in comparison with mainstream LLMs, HeCoNER (Llama3-8B) achieves the most substantial performance gains on the GENIA dataset in the biomedical domain, which features relatively complex nested structures. This further confirms that the HFAD module effectively bridges the "Semantic-Structural Divergence" between discriminative and generative models. By dynamically filtering noise during feature projection, it ensures that the model can inject high signal-to-noise ratio structural priors into the LLM across different parameter scales, thereby realizing robust and generalizable nested entity recognition.

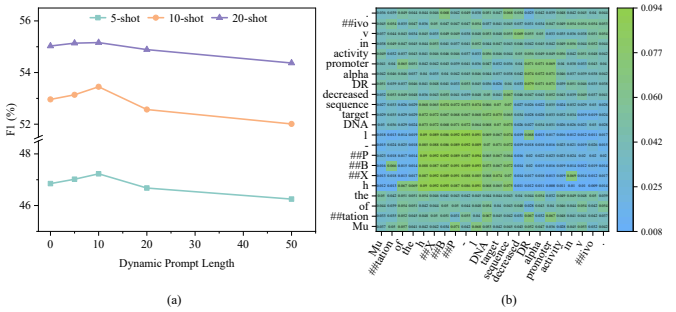
D. Ablation Study

As shown in Table II, given the representative nature of the complex nested structures in GENIA, we choose to present the ablation analysis performed on the GENIA dataset using HeCoNER (Qwen3-0.6B). Removing Instance-Adaptive Constraint Injection (w/o IACI) and Dynamic Prompt (w/o Dynamic Prompt) both resulted in performance degradation, preliminarily verifying the necessity of each component. The most significant performance deterioration occurred after replacing the HFAD module: simple linear projection (Linear) caused the 10-shot F_1 score to plummet by 9.87%, and even employing a Multi-Layer Perceptron (MLP) (with layers adjusted to maintain the same parameter magnitude as HFAD) still resulted in a 6.02% gap. This macro-level result indicates that the alignment of few-shot heterogeneous features requires interaction mechanisms more complex than static mapping. In the following sections, we will deeply analyze the working mechanisms of each module from different dimensions.



(a) Standard BERT (b) Dynamic Prompt-Guided BERT

Fig. 3. Impact of Dynamic Prompt on BERT feature distributions (t-SNE). (a) Feature visualization without Dynamic Prompt. (b) Feature visualization with Dynamic Prompt. Red and gray points denote entity and non-entity tokens, respectively.



(a) (b)

Fig. 4. (a): Impact of Dynamic Prompt Length on F_1 scores on the GENIA dataset. (b): Visualization of cross-attention weights in the HFAD module for a representative sample from the GENIA dataset. The axis labels correspond to the input text tokenized by BERT. Note that the prefix "##" denotes a subword unit, indicating that the token is a suffix belonging to the preceding word.

E. Detailed Analysis

1) Can the Dynamic Prompt in the DPGBE Module Effectively Mitigate Overfitting and Capture Latent Structural Features? To validate the core advantages of Dynamic Prompts in addressing the overfitting problem of BERT under few-shot conditions and extracting latent structural features, we conducted an in-depth verification from three dimensions: stability, feature distribution, and parameter sensitivity.

Stability in Extreme Low-Resource Scenarios: First, we compare different encoder training strategies under the 1-shot setting (Table III). Although Full Fine-tuning theoretically offers the largest optimization space, it exhibits extremely high instability on GENIA (std $\pm$11.25), indicating severe overfitting caused by single-sample updates. Even LoRA fine-tuning fails to eliminate this fluctuation. In contrast, our method (Frozen Encoder + Dynamic Prompt) demonstrates superior robustness, achieving the highest F1 (44.71) and the lowest standard deviation ($\pm$4.97). This proves that locking the general feature space via parameter freezing, while relying solely on prompt vectors for soft guidance, is the most reliable strategy for mitigating overfitting risks.

Visualization of Structural Activation: To investigate the underlying mechanism, we performed a t-SNE visualization analysis using BERT output vectors from 100 randomly selected samples of the GENIA test set (Fig. 3(a-b)). In the standard BERT feature space (Fig. 3(a)), the two classes of tokens appear highly intermixed and entangled. Conversely, the introduction of dynamic prompts (Fig. 3(b)) restructures the feature distribution: despite partial overlap at the boundary, the tokens exhibit a distinct trend towards clustering separation. This transition from disordered entanglement" to ordered separation" confirms that dynamic prompts effectively guide the frozen BERT to focus on syntactic structures and boundary cues, thereby yielding more discriminative representations.

Sensitivity Analysis on Dynamic Prompt Length: Fig. 4(a) illustrates a consistent "rise-then-fall" trend across all shot settings. Specifically, performance peaks when the dynamic prompt length reaches 10, indicating that a moderate increase in length effectively enhances the model's representation capacity. However, further extending the length (e.g., to 20 or 50) leads to a significant decline—sometimes even dropping below the baseline—suggesting that excessively long prompts introduce noise in data-scarce scenarios. Consequently, we fix the length at 10 to achieve an optimal balance between expressiveness and robustness.

2) Can HFAD Eliminate Noise and Improve the Recognition Capability for Deeply Nested Entities? We demonstrate the entity recognition performance across different entity nesting levels at various parameter scales and visualize the cross-attention weights, thereby conducting an in-depth verification of the denoising capability and mapping effectiveness of the HFAD module.

Visualization and Analysis of Attention Weights: To elucidate the heterogeneous feature space alignment mechanism of HFAD, we visualized the cross-attention weights (Fig. 4(b)). The heatmap reveals precise focusing: high weights are locked onto entity spans (e.g., "hXBP-1 DNA target sequence" and "DR alpha promoter"), with inner nested entities (e.g., "hXBP-1") receiving higher attention. Conversely, non-entity contexts (e.g., "activity in vivo") are assigned extremely low weights (blue regions). This confirms that HFAD dynamically suppresses non-entity semantic noise via the Query-Key mechanism and realizes high signal-to-noise ratio injection, effectively resolving the boundary overlap problem observed in Fig. 3.

Failure of Static Mapping (Noise Injection): As illustrated in Fig. 5, replacing HFAD with simple Linear Projection (Linear) or MLP not only fails to improve performance on complex nested entities (Level 2, 3+), but conversely leads to significant degradation in recognizing basic flat entities (Level 1), with performance falling even below the baseline LLM without information injection (w/o IACI). This phenomenon reveals the intrinsic defect of static mapping: in the absence of a dynamic filtering mechanism, simple dimensional alignment forcibly injects redundant features from the encoder as noise into the LLM, thereby compromising the model's intrinsic capability to recognize simple flat entities. Revisiting the feature boundary overlap observed in the earlier t-SNE results, direct static mapping inevitably propagates this semantic confusion, thus

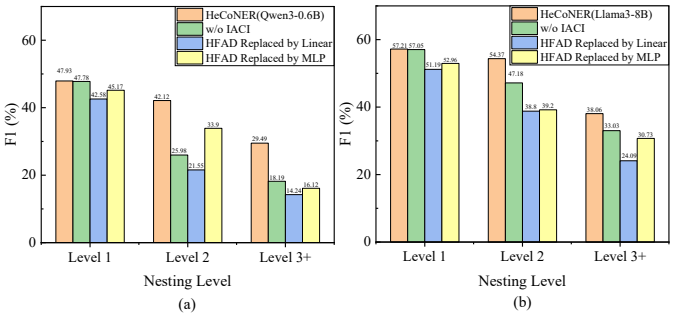


Fig. 5. Performance comparison of ablation variants across nesting levels: Level 1 (flat entities), Level 2 (one layer of nesting), and Level 3+. "w/o IACI" refers to the removal of all modules, utilizing only the LoRA-tuned LLM.

qualitatively explaining why linear layers introduce substantial noise

Quantitative Advantage of HFAD: In contrast, HFAD exhibits superior feature selection capabilities. From a numerical perspective, HFAD maintains a Level 1 F1 score comparable to w/o IACI, indicating effective noise filtering; meanwhile, it substantially outperforms all variants in complex nested scenarios (Level 2/3+). This performance of "zero loss" on flat entities and "high gain" on nested entities strongly proves that HFAD is an effective approach for heterogeneous feature alignment.

3) Can IACI Achieve Robust Constrained Decoding in Complex Nested Structures? Finally, we delve into whether the IACI strategy for the LLM is effective when handling complex entity nested structures. From Fig. 5, it can be clearly observed that when the IACI strategy is not used (w/o IACI), the model exhibits a typical "long-tail decay" effect: as the entity nesting depth increases (Level 1 $\rightarrow$ 3+), its F1 score experiences a sharp decline. This phenomenon confirms that in the absence of explicit structural constraints, standard instruction fine-tuning struggles to handle the recognition of complex nested entities. The introduction of IACI fundamentally changes this situation, and its corresponding performance improvement (particularly at Level 3+) validates our motivation for "Constrained Decoding."

F. Case Study

To intuitively verify the effectiveness of our method in mitigating boundary shift and hallucination in LLMs for few-shot nested NER tasks, we selected two representative cases from the GENIA dataset under the 5-shot setting. As shown in Fig. 6, comparing the baseline (Qwen3-0.6B) with HeCoNER not only illustrates specific error patterns but also highlights the correction capability of our method. In the first case (Test Sample 1), the baseline model failed to correctly recognize the nested entity "lymphocytes", identifying only the outer entity "B lymphocytes". More critically, it generated an entity "IgG" that was completely absent from the original text. This Hallucination phenomenon indicates that, in the absence of constraints, LLMs tend to over-rely on parameterized knowledge acquired during the pre-training phase (i.e., the

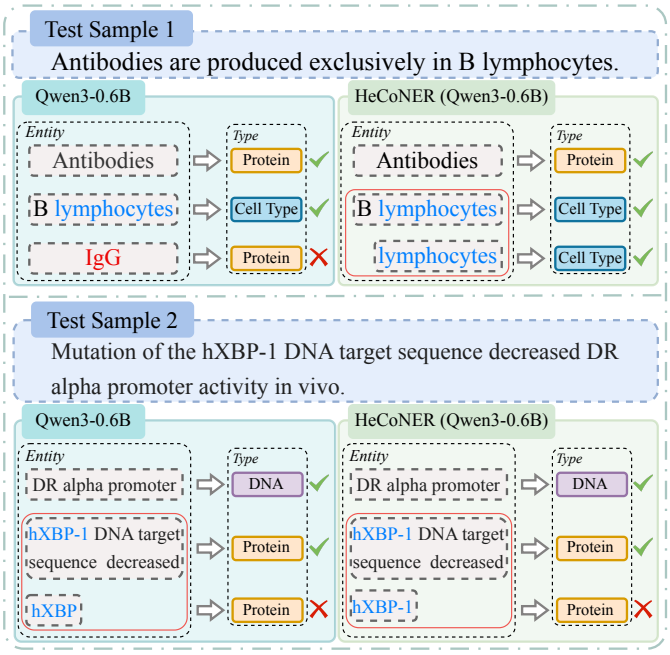


Fig. 6. case study on GENIA.

biological common sense that B lymphocytes produce IgG antibodies), thereby ignoring the actual constraints of the current context and leading to over-reasoning. In contrast, the HeCoNER method successfully suppressed this semantic hallucination and accurately extracted the entities actually present in the text. In the second case (Test Sample 2), although the baseline model did not produce hallucinations, it exhibited a typical Entity Boundary Shift problem. It erroneously truncated the correct entity "hXBP-1" to "hXBP". Although the two are highly correlated in medical semantics, the NER task requires strict span matching, and such subtle boundary deviations severely compromise system precision. Conversely, the HeCoNER method, benefiting from the entity structural features extracted by DPGBE, accurately locked onto the complete entity boundaries including the numerical suffix. These two cases validate the challenges of hallucination and boundary shift faced by LLMs in few-shot NER tasks and demonstrate the effectiveness of the instance-adaptive constraints introduced by HeCoNER in resolving these issues.

V. Conclusion

This paper presents HeCoNER, a heterogeneous collaborative framework for Few-Shot Nested NER. The framework innovatively integrates the feature extraction capabilities of discriminative models with the semantic reasoning advantages of Generative LLMs. Specifically, we validated the effectiveness of the Dynamic Prompts strategy, demonstrating that it can robustly capture entity boundary features even with frozen parameters. Furthermore, through the HFAD module and the IACI strategy, we successfully mitigated the noise issue across cross-model feature spaces and effectively com-

pressed the next-token search space of the LLM, achieving a transformation from open-ended generation to constrained decoding. The significant performance improvements across multiple benchmark datasets fully substantiate the robustness of HeCoNER in handling complex nested structures and also provide a new solution avenue for structured extraction tasks in low-resource scenarios.

REFERENCES

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticæ Investigationes*, vol. 30, no. 1, pp. 3–26, 2007. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad

[2] J. R. Finkel and C. D. Manning, "Nested named entity recognition," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 141–150.

[3] Y. Xu, Z. Yang, L. Zhang, D. Zhou, T. Wu, and R. Zhou, "Focusing, bridging and prompting for few-shot nested named entity recognition," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2621–2637. [Online]. Available: https://aclanthology.org/2023.findings-acl.164/

[4] M. Zhang, B. Wang, H. Fei, and M. Zhang, "In-context learning for few-shot nested named entity recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 026–10 030.

[5] Y. Xu, L. Zhang, and D. Zhou, "TECA: A two-stage approach with controllable attention soft prompt for few-shot nested named entity recognition," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 15 698–15 710. [Online]. Available: https://aclanthology.org/2024.lrec-main.1363/

[6] S. Zhu, Y. Yuan, L. Shi, and S. Xu, "Advancements in few-shot nested named entity recognition: The efficacy of meta-learning convolutional approaches," *Neurocomput.*, vol. 635, no. C, Jun. 2025. [Online]. Available: https://doi.org/10.1016/j.neucom.2025.129893

[7] H. Ming, J. Yang, S. Liu, L. Jiang, and N. An, "Mitigating prototype shift: Few-shot nested named entity recognition with prototype-attention contrastive learning," *Expert Syst. Appl.*, vol. 268, no. C, Apr. 2025. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.126293

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[9] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting few-sample bert fine-tuning," 2021. [Online]. Available: https://arxiv.org/abs/2006.05987

[10] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines," 2021. [Online]. Available: https://arxiv.org/abs/2006.04884

[11] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019. [Online]. Available: https://inria.hal.science/hal-02131630

[12] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022.

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[14] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Jan. 2025. [Online]. Available: https://doi.org/10.1145/3703155

[15] R. Han, C. Yang, T. Peng, P. Tiwari, X. Wan, L. Liu, and B. Wang, "An empirical study on information extraction using large language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.14450

[16] Y. Cai, Q. Liu, Y. Gan, R. Lin, C. Li, X. Liu, D. Luo, and J. JiayeYang, "Difinet: Boundary-aware semantic differentiation and filtration network for nested named entity recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 6455–6471.

[17] B. Wang and W. Lu, "Neural segmental hypergraphs for overlapping mention recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 204–214. [Online]. Available: https://aclanthology.org/D18-1019/

[18] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6470–6476.

[19] M. Rojas, F. Bravo-Marquez, and J. Dunstan, "Simple yet powerful: An overlooked architecture for nested named entity recognition," in *Proceedings of the 29th international conference on computational linguistics*, 2022, pp. 2108–2117.

[20] Y. Wang, H. Shindo, Y. Matsumoto, and T. Watanabe, "Nested named entity recognition via explicitly excluding the influence of the best path," *Journal of Natural Language Processing*, vol. 29, no. 1, pp. 23–52, 2022.

[21] P. Wang, R. Xu, T. Liu, Q. Zhou, Y. Cao, B. Chang, and Z. Sui, "An enhanced span-based decomposition method for few-shot sequence labeling," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5012–5024. [Online]. Available: https://aclanthology.org/2022.naacl-main.369/

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[23] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[24] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, and X. Qiu, "A unified generative framework for various NER subtasks," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 5808–5822. [Online]. Available: https://aclanthology.org/2021.acl-long.451/

[25] J. Chen, Q. Liu, H. Lin, X. Han, and L. Sun, "Few-shot named entity recognition with self-describing networks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5711–5722. [Online]. Available: https://aclanthology.org/2022.acl-long.392/

[26] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation." in *Lrec*, vol. 2. Lisbon, 2004, pp. 837–840.

[27] C. Walker, S. Strassel, J. Medero, and K. Maeda, "Ace 2005 multilingual training corpus-linguistic data consortium," *URL: https://catalog. ldc. upenn. edu/LDC2006T06*, 2005.

[28] T. Ohta, Y. Tateisi, J.-D. Kim, H. Mima, and J. Tsujii, "The genia corpus: An annotated research abstract corpus in molecular biology domain," in *Proceedings of the human language technology conference*. Morgan Kaufmann Publishers Inc. San Francisco, 2002, pp. 73–77.

[29] R. Ma, X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, and X. Huang, "Template-free prompt tuning for few-shot NER," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5721–5732. [Online]. Available: https://aclanthology.org/2022.naacl-main.420/

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.