DocIE@XLLM25: UIEPrompter: A Unified Training-Free Framework for universal document-level information extraction via Structured Prompt

Chengfeng Qiu[†], **Lifeng Zhou**[†], **Kaifeng Wei**, **Yuke Li**[*]
NetEase YiDun AI Lab, Hangzhou, China
{qiuchengfeng,hzzhoulifeng,hzweikaifeng,liyuke}@corp.netease.com

Abstract

We introduce UIEPrompter, a unified, training-free framework that secures 1st place in the ACL 2025 shared competition on universal document-level information extraction. UIEPrompter effectively addresses both named entity recognition and relation extraction without the need for annotated data. Leveraging large language models, UIEPrompter establishes a zero-shot baseline through role-specific prompts, which are then refined via few-shot guidance and constrained output generation prompt to align with competition schemas. Additionally, by integrating outputs from several large language models, we reduce individual model biases, thereby improving overall performance. Evaluated on the competition evaluation dataset, UIEPrompter showcases outstanding performance in document-level information extraction, ultimately securing first place. The implementation code is available on GitHub.

1 Introduction

Information extraction (IE) (Jeong and Kim, 2022; Paolini et al., 2021; Fei et al., 2022; Li et al., 2023) serves as a critical bridge between unstructured text and structured knowledge representation, enabling the identification of entities, their semantic types, and inter-entity relationships in texts. In recent years, IE has garnered significant attention from both academia and industry. The latest ACL 2025 shared task on universal document-level information extraction (DocIE) presents a novel challenge that requires the extraction of named entities (Labusch et al., 2019; Vacareanu et al., 2024; Shi and Kimura, 2024; Wang et al., 2023) and their relations (Huang et al., 2021; Cabot and Navigli, 2021; Efeoglu and Paschke, 2024; Xue et al., 2024; Peng et al., 2024) from documents.

Conventional approaches adopt a fragmented pipeline, deploying separate models for named entity recognition (NER) and relation extraction (RE). While effective in narrow contexts, this paradigm faces two fundamental limitations: (1) a heavy reliance on annotated training data for optimizing both models, and (2) the risk of cascading error propagation between the NER and RE stages, where misidentified entities can lead to incorrect relation predictions downstream.

To overcome these constraints, we propose UIEPrompter as shown in Figure 1, a training-free framework that unifies NER and RE through meticulous prompt engineering and the generative prowess of large language models (LLMs) (Liu et al., 2024; Arrieta et al., 2025; Kuo et al., 2025). The key advantages of our framework are:

- **Training-free architecture**: UIEPrompter eliminates the need for task-specific training data or model fine-tuning. Unlike supervised approaches that rely on domain-specific annotations, our framework achieves competitive performance purely through prompt engineering, leveraging LLMs' inherent knowledge without parameter updates.
- **Unified IE framework**: By integrating NER and RE into a single LLM-based architecture, UIEPrompter bypasses error propagation inherent in cascaded NER→RE pipelines. This joint modeling approach ensures entity-relation consistency while reducing system complexity.
- **Domain-adaptive few-shot guidance**: UIEPrompter incorporates competition-specific examples to align outputs with the characteristics of the target domain, thereby enhancing overall performance.
- **Strong performance in document-level information extraction**: According to the of-

[1] † Equal contribution.

[2] ∗ Corresponding author.

Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 317–321
August 1, 2025 ©2025 Association for Computational Linguistics

ficial leaderboard, our architecture achieves the top overall score and secured first place in both the NER and RE tracks, with a significant lead over the second-place competitors.

2 Method

2.1 System Architecture

As shown in Figure 1, based on the input, we design a basic template that defines the LLM as an information extraction expert and specifies the task requirements. To further enhance the model's comprehension and generation capabilities, we introduce a few-shot guidance stage. During this phase, the model is provided with a small set of input-output examples that illustrate desired behaviors, enabling the model to better understand task objectives and expected generation patterns. Concurrently, we impose constraints on the content generation process to ensure outputs strictly adhere to predefined formats and satisfy task-specific requirements. After inputting these prompts into multiple large language models, we fuse the outputs to mitigate biases inherent in individual models, thereby enhancing the robustness and reliability of the final results. It is evident that UIEPrompter is a unified information extraction system based on LLMs. By integrating NER and RE into a single LLM architecture, UIEPrompter effectively circumvents the error propagation inherent in cascaded NER-to-RE pipelines. This joint modeling approach not only ensures consistency between entities and relations but also simplifies system complexity.

2.2 Basic Template

The Basic template encompasses role assignment and a clear task definition. It is structured around the following instruction:

```
You are an expert in document
NER and triplet extraction. I
will provide you with the domain
of the document, the document
text, a set of NER entity types,
and a set of relationship types
for triplet extraction. Please
help me extract the NER entities
and triplets that appear in the
document based on the input.
```

2.3 Few-shot Guidance

Recognizing that different individuals or models may exhibit variations in preferences for the same problem, and to align with the training set's preferences regarding officially recognized labels, we offer a case from the training set as a guiding example. The few-shot input and output example is as follows:

```
## Example 1
## The input is:
    "domain":    "...",   "doc":
"...",
    "NER_set":  [...],  "RE_set":
[...]
## The output is: ...
```

2.4 Constrained Output Generation

To ensure that the output from the large model can be parsed by the evaluation function, strict formatting constraints must be imposed. Additionally, to encourage the model to output a complete and agreed-upon format, we instruct it to omit any reasoning process. The prompts reflecting these requirements can be structured as follows:

```
## The output format must be
{"entities":"xxx","triples":"xxx"}.
## In the output, "entities"
represents the extracted NER
entities, while "triples"
represent the extracted triples.
## I do not need any analysis
or extraneous commentary; please
provide only the JSON formatted
result as specified!!
## Please ensure that the JSON is
valid and will not cause errors
when printed!!
```

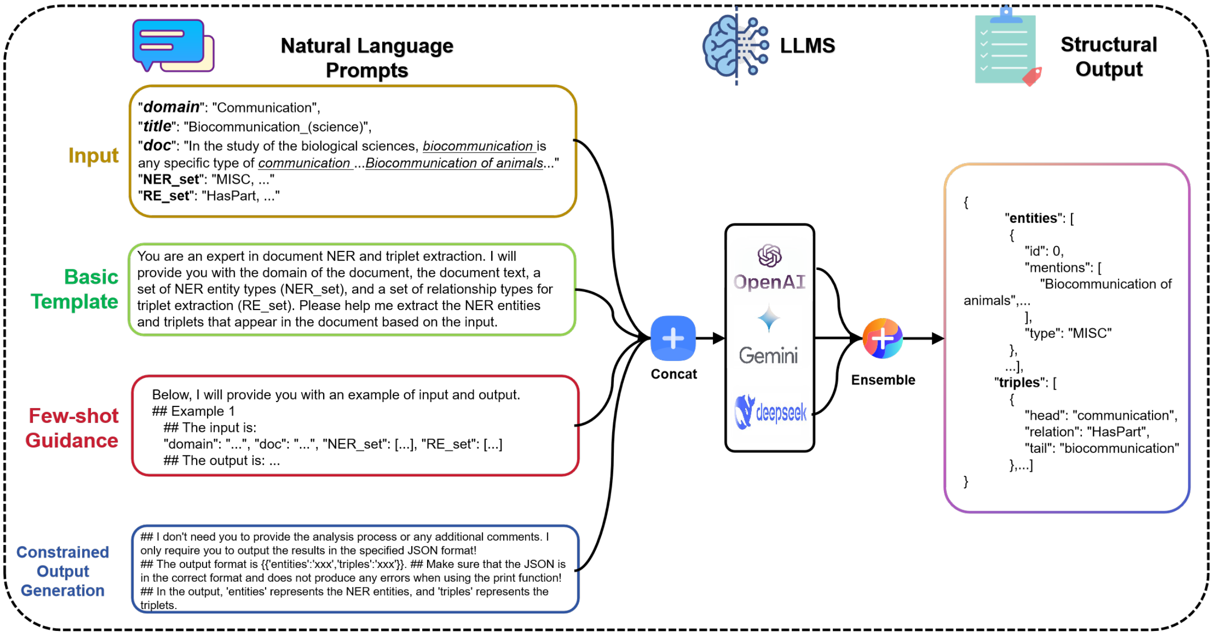
2.5 Ensemble to Boost Performance

To achieve better results, we input the prompt into multiple large language models and then combine their outputs. Our approach involves merging the outputs from different models and removing duplicates to obtain the final fused result. The experiment results demonstrate that this method is simple yet effective.

3 Experiments

3.1 Datasets

We utilize the DocIE development and evaluation datasets, which consist of 29 domain-specific



"*domain*": "Communication",
"*title*": "Biocommunication_(science)",
"*doc*": "In the study of the biological sciences, *biocommunication* is any specific type of *communication* ...*Biocommunication of animals*..."
"**NER_set**": "MISC, ..."
"**RE_set**": "HasPart, ..."

You are an expert in document NER and triplet extraction. I will provide you with the domain of the document, the document text, a set of NER entity types (NER_set), and a set of relationship types for triplet extraction (RE_set). Please help me extract the NER entities and triplets that appear in the document based on the input.

```
Below, I will provide you with an example of input and output.
## Example 1
  ## The input is:
  "domain": "...", "doc": "...", "NER_set": [...], "RE_set": [...]
  ## The output is: ...
```

I don't need you to provide the analysis process or any additional comments. I only require you to output the results in the specified JSON format!
The output format is {{'entities':'xxx','triples':'xxx'}}. ## Make sure that the JSON is in the correct format and does not produce any errors when using the print function!
In the output, 'entities' represents the NER entities, and 'triples' represents the triplets.

```
{
        "entities": [
         {
            "id": 0,
            "mentions": [
                "Biocommunication of
animals",...
            ],
            "type": "MISC"
         },
        ...],
      "triples": [
         {
            "head": "communication",
            "relation": "HasPart",
            "tail": "biocommunication"
         },...]
}
```

Figure 1: The system architecture of UIEPrompter.

Table 1: Competition leaderboard

Place	Participant	F1-EI↑	F1-EC↑	F1-REG↑	F1-RES↑	F1-AVG↑
1	**qqpprun(UIEPrompter)**	**65.52**	**32.20**	**5.40**	**5.11**	**27.06**
2	UIT-SHAMROCK	55.65	26.11	4.19	4.01	22.49
3	check_out	59.15	18.17	4.38	4.12	21.46
4	ScaDS.AI	32.86	16.19	3.29	3.01	13.84

datasets. We used the development dataset to identify our optimal settings, which were then applied as the final settings for evaluation on the evaluation datasets. More details can be found on the huggingface webcite (doc).

3.2 LLM Selection

In this study, we select three leading state-of-the-art large language models as our meta-models: OpenAI's o3-mini, Google's Gemini-2.0-flash, and DeepSeek's Deepseek-v3.

3.3 Evaluation Metric

We use the competition-specified metrics to evaluate the effectiveness of our models. For the NER track, the competition employs entity identification F1 (F1-EI) and entity classification F1 (F1-EC). In the RE track, the metrics include general mode F1 (F1-REG) and strict mode F1 (F1-RES). The final evaluation metric for the competition is the average of these four F1 scores, denoted as F1-AVG. Detailed definitions of these metrics can be found on the official website.

4 Experimental Results

4.1 Main results

The main results of UIEPrompter on the evaluation dataset are summarized in Table 1. We secure the championship title with an overall F1-AVG of 27.06%, surpassing the runner-up, who scored 22.49%, by nearly 5%. Notably, we also achieved first place in all four sub-F1 metrics: 65.52% for F1-EI and 32.20% for F1-EC in the named entities recognition task, and 5.40% for F1-REG and 5.11% for F1-RES in the relation extraction task. Based on the results, we can identify three key advantages of UIEPrompter:

- **Dominance in the named entities recognition task**: Our F1-EI of 65.52% and F1-EC of 32.20% significantly outperformed all competitors, showcasing our superior accuracy in NER task.

- **Strong performance in relation extraction task**: Despite lower absolute F1 scores in RE tasks, UIEPrompter achieves the highest scores in both F1-REG and F1-RES, indicating robust consistency in the task of relation extraction.

Table 2: Ablation study of the key components. BT stands for basic template. FSG refers to few-shot guidance, and COG denotes constrained output generation prompt.

Model	BT	FSG	COG	F1-EI↑	F1-EC↑	F1-REG↑	F1-RES↑	F1-AVG↑
o3-mini	✓			0	0	0	0	0
	✓	✓		32.82	15.06	9.29	5.25	15.61
	✓	✓	✓	32.39	14.34	16.23	10.53	18.37
Deepseek-v3	✓			0	0	0	0	0
	✓	✓		28.13	23.26	26.71	5.77	20.97
	✓	✓	✓	31.89	24.76	19.73	9.03	21.35
Gemini-2.0-flash	✓			0	0	0	0	0
	✓	✓		38.54	31.06	15.88	3.43	22.23
	✓	✓	✓	41.04	33.16	17.63	2.15	**23.50**

Table 3: Model ensemble results on the development dataset.

Gemini-2.0-flash	o3-mini	Deepseek-v3	F1-EI↑	F1-EC↑	F1-REG↑	F1-RES↑	F1-AVG↑
✓			41.04	33.16	17.63	2.15	23.50
✓	✓		44.47	28.09	22.59	7.97	**25.78**
✓	✓	✓	43.34	27.36	22.73	8.46	25.47

- **Balanced Competence**: The results demonstrate a strong performance across both NER and RE subtasks, showing no significant weaknesses when compared to other participants.

4.2 Ablation Study

To better illustrate the rationale behind our approach, we conducted an ablation study on the development dataset, with results presented in Table 2. It is important to note that the apparent "zero performance" observed with BT does not indicate the models' incapability, but rather reflects failures in the formatting parser. When structural format constraints are absent, outputs can become syntactically invalid (e.g., not JSON format or missing brackets). The results reveal that incorporating few-shot guidance, which provides the model with examples, effectively applies implicit output constraints and improves the performance of the large language model. Furthermore, additional enhancements can be achieved through the use of constrained output generation (COG) prompts, which impose explicit constraints.

Regrading architecture choices, our findings indicate that Gemini-2.0-flash achieved the best overall performance, reaching an F1-AVG of 23.50%. This superior performance can be attributed to its significantly higher F1-EI and F1-EC scores compared to other models, showcasing its strength in the NER task. However, it falls short of the other two models in the F1-REG and F1-RES metrics, suggesting relatively weak relation extraction capabilities. In other words, no single model performs optimally across both tasks simultaneously. To tackle this issue, we implemented a model ensemble approach to improve the overall performance of the framework.

The results of these models fusion are presented in Table 3. The data clearly indicates that the fusion of the Gemini-2.0-flash and o3-mini models achieved the highest overall performance, with an F1-AVG of 25.78% on the development dataset. Additionally, we observed that the fusion of three models on the validation set yielded a performance of 25.47%, which is quite close to the 25.78% achieved by the best combination. To evaluate performance differences on the evaluation set, we found that the fusion of the three models yielded the best results overall.

5 Conclusion

In this work, we present UIEPrompter, a unified and training-free framework that secured first place in the ACL 2025 shared competition on universal document-level information extraction. Our framework adeptly addresses the challenges of named entity recognition and relation extraction in document-level contexts without relying on annotated training data. By leveraging role-specific prompts for zero-shot initialization and adapting to competition schemas through few-shot guidance and constrained output generation prompt, UIEPrompter demonstrates remarkable information extraction performance and generalization capabilities.

References

shuyi-zsy/DocIE,datasets at hugging face. https://huggingface.co/datasets/shuyi-zsy/DocIE.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025. o3-mini vs deepseek-r1: Which one is safer? *arXiv preprint arXiv:2501.18438*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Sefika Efeoglu and Adrian Paschke. 2024. Retrieval-augmented generation-based relation extraction. *arXiv preprint arXiv:2404.13397*.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. *Advances in Neural Information Processing Systems*, 35:15460–15475.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. *arXiv preprint arXiv:2109.04901*.

Yuna Jeong and Eunhui Kim. 2022. Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, 10:60805–60813.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.

Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. 2019. Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th conference on natural language processing*, pages 9–11.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. Codeie: Large code generation models are better few-shot information extractors. *arXiv preprint arXiv:2305.05711*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. Metaie: Distilling a meta model from llm for all kinds of information extraction tasks. *arXiv preprint arXiv:2404.00457*.

Yuning Shi and Masaomi Kimura. 2024. Bert-based models with attention mechanism and lambda layer for biomedical named entity recognition. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, pages 536–544.

Robert Vacareanu, Enrique Noriega-Atala, Gus Hahn-Powell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2024. Active learning design choices for ner with transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 321–334.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: Document-level relation extraction with large language models. *arXiv preprint arXiv:2403.14888*.