

# MPL: Multiple Programming Languages with Large Language Models for Information Extraction

Bo Li<sup>1</sup>, Gexiang Fang<sup>2</sup>, Wei Ye<sup>2\*</sup>, Zhenghua Xu<sup>1\*</sup>, Jinglei Zhang<sup>2</sup>,

Hao Cheng<sup>3</sup>, Shikun Zhang<sup>2</sup>

<sup>1</sup>State Key Laboratory of Intelligent Power Distribution Equipment and System, School of Health Sciences and Biomedical Engineering, Hebei University of Technology

<sup>2</sup>National Engineering Research Center for Software Engineering, Peking University

<sup>3</sup>School of Artificial Intelligence, Hebei University of Technology  
deepblue.lb@gmail.com, wye@pku.edu.cn, zhenghua.xu@hebut.edu.cn

## Abstract

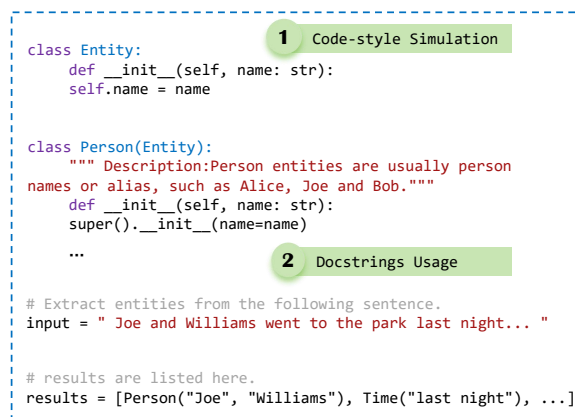
Recent research in information extraction (IE) focuses on utilizing code-style inputs to enhance structured output generation. The intuition behind this is that the programming languages (PLs) inherently exhibit greater structural organization than natural languages (NLs). This structural advantage makes PLs particularly suited for IE tasks. Nevertheless, existing research primarily focuses on Python for code-style simulation, overlooking the potential of other widely-used PLs (e.g., C++ and Java) during the supervised fine-tuning (SFT) phase. In this research, we propose **Multiple Programming Languages** with large language models for information extraction (abbreviated as **MPL**), a novel framework that explores the potential of incorporating different PLs in the SFT phase. Additionally, we introduce function-prompt with virtual running to simulate code-style inputs more effectively and efficiently. Experimental results on a wide range of datasets demonstrate the effectiveness of MPL. Furthermore, we conduct extensive experiments to provide a comprehensive analysis. We have released our code for future research <sup>1</sup>.

## 1 Introduction

Large language models (LLMs) (Brown et al., 2020; Hoffmann et al., 2022; Smith et al., 2022; Ouyang et al., 2022; Touvron et al., 2023) trained on massive natural language have demonstrated their superiority in a wide range of natural language processing tasks, such as machine translation (Li et al., 2023f; Zhang et al., 2023) and dialogue systems (Yang et al., 2024b; Nananukul and Wongkamjan, 2024). Nevertheless, researchers have found that LLMs do not achieve desirable performance on information extraction tasks, such as named entity recognition and relation extraction

\*Co-corresponding authors: wye@pku.edu.cn, zhenghua.xu@hebut.edu.cn

<sup>1</sup><https://github.com/PKU-Fgx/MPL>



```
class Entity:
    def __init__(self, name: str):
        self.name = name

class Person(Entity):
    """ Description: Person entities are usually person
    names or alias, such as Alice, Joe and Bob. """
    def __init__(self, name: str):
        super().__init__(name=name)
    ...

# Extract entities from the following sentence.
input = " Joe and Williams went to the park last night... "

# results are listed here.
results = [Person("Joe", "Williams"), Time("last night"), ...]
```

Figure 1: The typical procedure of code-style information extraction system, which mainly contains two components: the code-style simulation and the docstring usage.

(Li et al., 2023a; Xu et al., 2023; Han et al., 2023). Since IE requires extracting structured knowledge from complex and variably expressed natural language, relying solely on natural language instructions, even with tailored prompts, may not be sufficient for generating structured outputs effectively (Li et al., 2023d; Guo et al., 2023; Li et al., 2024c).

Recent studies have increasingly focused on exploring code-style input for IE tasks, as programming languages (PLs) are more formal and structured than natural languages (NLs) (Van Roy and Haridi, 2004; Dunn, 2019), making them better suited for structured generation tasks. Broadly speaking, existing works (Wang et al., 2022b; Guo et al., 2023; Li et al., 2023d; Sainz et al., 2023; Bi et al., 2024; Li et al., 2024c) mainly involve two components: **1) Code-style Simulation**—where the textual input is reformulated into a code-style representation using approach like `class-prompt`, which includes base `class` and `subclass` to represent the main task and fine-grained labels, along with additional elements to maintain standard code formatting. **2) Docstring Usage**—which involves

using docstrings or guidelines for two main purposes: adding task and label descriptions within each class definition and providing explicit instructions for the task’s inputs and outputs. The above procedures are briefly shown in Figure 1. Overall, existing works formulate IE tasks as code-style inputs, enabling LLMs to generate structured outputs more effectively and achieve strong performance in IE tasks.

Although `class-prompt` is commonly used to reformulate a given IE task into code-style input, it often requires defining numerous repetitive elements to maintain a standard format, such as the constructor `__init__` and the attribute assignment `self.name = name`. Besides, some label descriptions are repeated multiple times in docstrings due to the structural requirements. These redundancies not only result in verbose code but also significantly increase input length, negatively impacting both performance and training efficiency. To address these challenges, we propose `function-prompt` with a virtual running component, a more lightweight approach that wraps the task definitions, label descriptions and textual input into a single function, allowing LLMs generate the outputs by virtually executing the above function.

Additionally, previous work on code-style IE systems has primarily focused on Python, overlooking the multilingual nature of programming languages. Specifically, a wide range of commonly used PLs possess distinct inherent features, that could make them more suitable for specific IE tasks. For instance, Python’s clean and readable syntax (Sarkar, 2016; Holst and Dobslaw, 2021) facilitates handling complex nested structures and enhances clarity, making it a standard choice for code-style in IE systems. However, Python’s dynamic typing can lead to ambiguous types, which may introduce errors that are only detected during runtime. In contrast, C++ enforces strong typing, requiring explicit declarations of variable types (Stroustrup, 2013, 2014), which helps prevent type-related errors early in development. Additionally, Java’s object-oriented nature (Blanchet, 1999; Poo et al., 2008) simplifies the representation of complex data structures, demonstrating the advantages of considering a variety of languages for diverse IE needs. In this paper, we propose **Multiple Programming Languages** with large language models for information extraction (shortened as **MPL**), a novel perspective that explores the integration of different

PLs in the code-style IE system. To be specific, for a given textual input, we use three programming languages to simulate the coding process, i.e., Python, C++ and Java. These PLs are chosen for their common use and distinct characteristics, and the code-style inputs will be used in the supervised fine-tuning (SFT) process upon LLMs, aiming at exploring the cooperation between different PLs when dealing with IE tasks.

We conduct comprehensive experiments across a variety of IE tasks to evaluate the effectiveness of MPL. Specifically, we use widely used IE datasets for named entity recognition, relation extraction and event extraction to create datasets for SFT in the code-style format. We find that with the help of the simple and efficient `function-prompt`, MPL achieves significantly better performance compared to previous works that relied on larger model architectures or extensive external pre-training datasets. Additionally, our extensive experimental analysis of code-style simulations, LLM usage, and complementarity between different PLs provides valuable insights, serving as practical guidelines for future research. To sum up, our main contributions are as follows:

- We are the first to utilize multiple programming languages with LLMs for IE tasks, aiming to explore the potential of combining different PLs. MPL’s strategy involves generating various code-style inputs from a single textual input and performing SFT on LLMs. Additionally, we developed the `function-prompt` with virtual running for more efficient and effective code-style simulations.
- We carry out thorough experiments and analysis using diverse IE datasets, where MPL consistently achieves strong outcomes. Our experiments also extend to exploring different aspects of code-style simulation, LLM selection, and the interaction between various PLs, providing valuable insights for ongoing research.

## 2 Proposed Method

In this paper, we propose **MPL**, a new framework that leverages multiple programming languages with large language models for information extraction. First, we briefly introduce the commonly used `class-prompt`, and then we propose a new

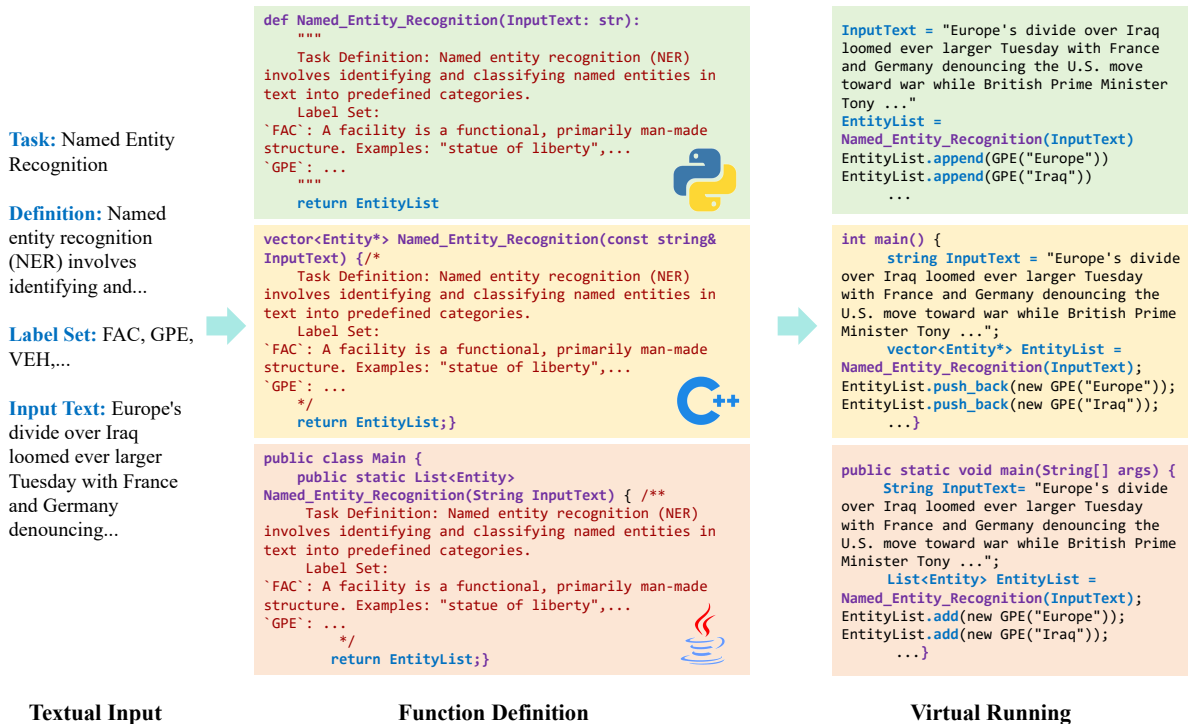


Figure 2: Our framework utilizes multiple programming languages, i.e, Python, C++, and Java, to convert elements from IE tasks and target textual inputs into code-style formats. To enhance the simulation process and help LLMs in processing textual inputs and generating outputs more naturally, we introduce the function-prompt with function definition and virtual running components. Better viewed in color.

design named function-prompt, which simulates the coding process in a more lightweight and effective way. Next, we introduce how to transform textual input into code-style representations across multiple PLs. Finally, we provide training details to clarify the implementation.

## 2.1 Class-prompt

The Class-prompt comprises two hierarchical levels: base class and subclass, representing the label definitions of a given IE task at different granularities. An overview is shown in Figure 1. The Base class serves as the parent class of all fine-grained entity, relation and event labels, denoted as Entity, Relation and Event. For instance, all entity labels are grouped under the Entity base class. Typically, the base class contains a constructor `__init__` and the attribute assignment `self.name = name`. The subclass utilizes class inheritance to represent more specific labels, such as a PERSON entity type being a subclass of Entity, defined as `class PERSON(Entity)`. Some approaches also incorporate event triggers within the corresponding event subclass when applicable. Besides, docstrings or guidelines are often used within subclass to wrap the task

definitions and label descriptions, providing necessary information for understanding the given IE task and fine-grained labels easily and accurately.

## 2.2 Function-prompt

Although class-prompt effectively reformulates IE tasks into a standard code-style representations, it requires defining numerous repetitive elements to maintain the standard code format, and label descriptions often need to be repeated multiple times in docstrings or guidelines due to the nature of programming languages. These drawbacks significantly increase the length of the code-style inputs, negatively impacting both performance and training efficiency. To address these issues, this paper introduces function-prompt, designed to simplify the coding process while preserving the structural benefits of code-style input. The overall framework is shown in Figure 2. Specifically, function-prompt mainly involves the following two parts:

**Function Definition.** Taking NER as an example, we first define a function that takes `InputText` as an input variable. Next, we add the task definition and label descriptions in the docstring, providing crucial information for un-

derstanding the given task and candidate labels. Specifically, we use “Task Definition” and “Label Set” to clearly highlight the target task and its corresponding candidate labels. At the end of the function, a return statement outputs the prediction, making our code-style input both complete and more reflective of actual coding practices. Besides, compared to the class-prompt, our function-prompt eliminates the need for repetitive additional elements and label descriptions, thereby providing a simpler and more lightweight method for simulating code-style inputs.

**Virtual Running.** Unlike previous approaches that solely rely on docstrings to guide models in processing inputs and generating outputs—a method that is somewhat unnatural for code-style inputs—our function-prompt incorporates a virtual running step. This step involves taking the textual input as input to the function and executing it virtually. Virtual running not only makes interaction with LLMs more intuitive under the code-style representation but also enhances their ability to generate structured outputs efficiently. During the generation phase, outputs are stored and displayed using a list. For instance, with the NER task, the `Named_Entity_Recognition` function runs virtually, and each identified entity is subsequently appended to an `EntityList`.

### 2.3 Multiple Programming Languages for Information Extraction

Instead of relying solely on Python, this study employs a broader range of programming languages, including Python, C++, and Java, to simulate the code-style inputs. MPL represents a given textual input from multiple programming languages perspectives and learns complementary information. For a given task and a textual input, we show different code-style inputs in Figure 2, and the whole input for each programming language is provided in the Appendix C.

To convert a given IE task and textual input into a code-style representation, we develop transforming functions for each programming language. These functions reformulate the textual input to align with the syntax and structure of the respective programming languages. In this process, we follow two key guidelines: 1) ensuring that all necessary information, such as the task name and definition, candidate labels and descriptions, along with the virtual running component, is wrapped within the chosen programming language; 2) maintaining consistency

in the logic of the code design and the execution flow across Python, C++, and Java. Following these principles, we generate code-style inputs that convey similar semantic meanings while adopting diverse formats for IE tasks.

### 2.4 Training and Testing

During the training phase, these code-style inputs are used in the supervised fine-tuning (SFT) phase for IE tasks. We apply the standard Next Token Prediction loss, training our models solely on the actual output tokens while excluding the input sequences. For example, in the NER task shown in Figure 2, we compute the loss only after `EntityList =`, and the output order follows the label definition order. All the code-style inputs with different programming languages are randomly shuffled before feeding into the model. During the testing phase, unless specified otherwise, inputs from three programming languages are used to derive predictions, which are then aggregated through a voting mechanism to determine the final results <sup>2</sup> (referred to as the default setting).

## 3 Experimental Setup

### 3.1 Dataset

In this paper, we use the following widely used datasets across various IE tasks. These tasks include Named Entity Recognition (NER), Relation Extraction (RE), Event Argument Extraction (EAE) and Event Extraction (EE). Specifically, we use ACE05-NER<sup>3</sup>, BC5CDR (Wei et al., 2016), DIANN (Fabregat et al., 2018), NCBID (Dogan and Lu, 2012), WNUT2017 (Derczynski et al., 2017) CoNLL03 (Sang and De Meulder, 2003), and OntoNotes 5.0 (Pradhan et al., 2013) for NER, ACE05-RE for RE. As for event-related tasks, we use ACE05-EAE and RAMS (Ebner et al., 2019) for EAE, and ACE05-EE for EE. The above datasets cover a wide range of topics and are suitable for the SFT phase. Consistent with previous works (Sainz et al., 2023; Li et al., 2024c), the Micro-F1 score is used to evaluate performances. The star (\*) is used to indicate that the OntoNotes 5 dataset was aligned to the GOLLIE (Sainz et al., 2023) setup, which differs significantly from that used by KnowCoder in both training and testing.

<sup>2</sup>We also evaluate the model using a single PL during testing and observed only a slight performance degradation, please refer Section 4.3 for more details.

<sup>3</sup><https://catalog.ldc.upenn.edu/LDC2006T06>



	CodeLLaMA-7B		CodeLLaMA-13B					
	GoLLIE	MPL	GoLLIE	MPL	SLM	KnowCoder	GoLLIE-34B	MPL-8B
<b>ACE05-NER</b>	88.1	89.6	89.4	90.6	86.6	86.1	89.6	91.4
<b>BC5CDR</b>	87.5	87.9	87.9	87.8	91.9	89.3	88.4	89.6
<b>CoNLL03</b>	92.8	93.1	93.0	93.2	93.0	95.1	93.1	93.5
<b>DIANN</b>	79.4	84.8	82.6	85.7	74.8	94.7	84.1	85.4
<b>NCBID</b>	85.4	87.2	86.5	88.4	90.2	83.8	85.8	88.1
<b>OntoNotes5*</b>	83.4	83.7	84.0	84.7	84.6	-	84.6	85.7
<b>WNUT2017</b>	52.0	53.1	50.5	52.6	60.2	66.4	54.3	52.6
<b>ACE05-RE</b>	63.6	65.8	67.5	66.0	66.1	64.5	70.1	70.8
<b>ACE05-EAE</b>	66.0	71.2	67.8	72.9	54.8	70.3	68.6	72.8
<b>RAMS</b>	48.7	48.9	49.6	49.4	48.6	-	51.2	50.9
<b>ACE05-EE</b>	72.2	70.7	70.9	70.1	73.4	74.2	71.9	72.7
<b>Avg.Score</b>	74.4	<b>76.1(+1.7)</b>	75.4	<b>76.5(+1.1)</b>	74.9	-	76.5	<b>77.6(+1.1)</b>

Table 1: We conducted extensive experiments on CodeLLaMA-7B, CodeLLaMA-13B and LLaMA3-8B, where MPL-8B denotes the backbone is LLaMA3-8B. The values in brackets represent the performance differences between our MPL approach and the corresponding GoLLIE model. To minimize randomness, we tested each model five times and reported the average performance. Additionally, we conducted t-tests to compare our results with previous results, confirming that our results are statistically significant with a  $p$ -value of less than 0.05.

### 3.2 Training Details

We use the base version of StarCoder-v2 (Lozhkov et al., 2024), LLaMa2-7B (Touvron et al., 2023), CodeLLaMA-7B, CodeLLaMA-13B (Roziere et al., 2023) and LLaMA3-8B<sup>4</sup> in our experiments<sup>5</sup>. We use QLoRA (Dettmers et al., 2024) with  $\text{lora\_rank} = 96$  and  $\text{lora\_alpha} = 192$  for efficient training, the dropout ratio is set to 0.1. The models were trained with batch size of 8 and a learning rate of  $1e-4$  with a cosine scheduler and 0.05 warmup. The maximum length of input is 2048, the training epoch is 5.

### 3.3 Compared Models

We mainly compare our method with GoLLIE (Sainz et al., 2023) and KnowCoder (Li et al., 2024c), both of which have recently achieved SoTA results and involve supervised fine-tuning phase on LLMs. GoLLIE uses specific data processing and guideline expansion operations, while KnowCoder uses 33 specific domain IE datasets in SFT stage. Besides, KnowCoder includes a schema understanding phase that contains millions of instances across various external knowledge to further pre-train LLMs. We also compare MPL with previous

data-specific supervised fine-tuning, which uses small language models as backbone networks, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), denoted as SLM. We present these reference models in the Appendix B.

Furthermore, we have designed several model variants to conduct ablation studies. Specifically, to determine whether simply increasing data size can yield significant improvements, we trained a model with a  $3\times$  larger dataset based on python-style inputs, denoted as Python $_{3\times}$ . We trained a model named as MPL<sub>sampled</sub>, which randomly selects one PL instance per input, resulting in a dataset one-third the size of the full multiple PL dataset. MPL<sub>sampled</sub> aims to investigate if employing the same data size in multiple PL settings can achieve satisfactory performance. Additionally, we report the performance of models trained with a single programming language during the SFT phase for a detailed analysis.

## 4 Results and Analysis

We conducted extensive experiments to verify the effectiveness of MPL, below are our results and findings.

### 4.1 Main Results

The main results are shown in Table 1, from which we can see that: 1) MPL achieves the best overall performance compared to both SLM-based and

<sup>4</sup><https://ai.meta.com/blog/meta-llama-3/>

<sup>5</sup>We also use Qwen2.5-7B (Yang et al., 2024a) and QwenCoder2.5-7B (Hui et al., 2024) to further test MPL’s generalization ability with different LLM backbones, and the results are reported in Appendix E.

	<i>Average Score</i>		<i>Language Capability</i>	
	<i>Python</i>	<i>MPL</i>	<i>MMLU (NL)</i>	<i>HumanEval Pass@1 (PL)</i>
<b>StarCoder-v2</b>	72.7	73.8	38.3	35.4
<b>LLaMa2-7B</b>	75.9	76.6	45.3	12.2
<b>LLaMA3-8B</b>	76.4	77.6	66.7	37.2

Table 2: The results of various backbone LLMs trained with Python and MPL. We also report the the *natural language ability* (shortened as *NL*) and *programming language ability* (shortened as *PL*) of each backbone LLMs. These results show that the natural language ability is more important than programming language ability in the case of information extraction tasks.

LLM-based SoTA models. Notably, MPL-8B outperforms the largest GoLLIE model (34B parameters) by 1.1% and surpasses the previous SLM-based SoTA by 2.7%. Besides, MPL demonstrates competitive performance against KnowCoder, despite KnowCoder incorporates a schema understanding phase with millions of instances and uses 33 domain-specific IE datasets in the SFT stage. 2) For the same backbone, MPL consistently outperforms GoLLIE, achieving a 1.7% improvement for CodeLLaMA-7B and 1.1% for CodeLLaMA-13B. Moreover, MPL trained with CodeLLaMA-13B even achieves performance comparable to GoLLIE trained with the CodeLLaMA-34B model. These findings verify strength of MPL across various IE tasks, supporting both the motivation behind our research and the effectiveness of our model. 3) MPL consistently improves across different model configurations, including CodeLLaMA-7B, CodeLLaMA-13B and LLaMA3-8B. This suggests that MPL generalizes well across various model sizes. Its ability to deliver strong results with different backbones highlights its robustness and versatility for information extraction tasks.

<b>Dataset</b>	<b>GoLLIE-34B</b>	<b>MPL</b>
WikiEvents-NER	81.3	<b>82.1</b>
Broad_Twitter	50.3	<b>55.5</b>
MIT_Movie	<b>62.4</b>	62.2
WikiEvents-EAE	<b>50.7</b>	47.7

Table 3: Comparison of GoLLIE-34B and MPL across various zero-shot IE datasets.

## 4.2 Zero-shot Performance

To assess the generalization capability of MPL beyond its training distribution, we evaluate it on four information extraction (IE) benchmarks that

are excluded from the training set: WikiEvents-NER (Li et al., 2021), Broad\_Twitter (Derczynski et al., 2016), MIT\_Movie (Liu et al., 2013) and WikiEvents-EAE (Li et al., 2021). The results of MPL-8B and GoLLIE-34B are presented in Table 3. On WikiEvents-NER, MPL achieves the highest score, despite having fewer parameters and undergoing lighter training. Even on datasets where GoLLIE-34B performs slightly better, the performance gap remains minimal. These findings indicate that MPL exhibits strong generalization across diverse datasets and task structures.

## 4.3 Ablation Study

To evaluate the effectiveness of proposed settings in MPL, we conducted a thorough ablation study, with results shown in Table 4. **1) Single vs. Multiple Programming Languages:** In the first block, we trained LLaMA3-8B model using a single PL with the function-prompt. The results reveal that performance with a single PL significantly lags behind MPL, highlighting the advantage of using multiple PLs for IE tasks. **2) Data Construction Methods:** In the second block, we explored different data construction methods. When we increased the Python-style dataset size threefold (Python<sub>3×</sub>), we observed severe overfitting and a notable drop in performance compared to MPL (-1.5% on the average score). In contrast, randomly sampling one-third of the MPL training data resulted in a surprisingly competitive performance (-0.5% decrement). These findings suggest that MPL’s improvements stem from using multiple PLs with diverse representations, rather than merely increasing the dataset size. **3) Virtual Running Component:** In the final block, we evaluated the virtual running component, which simulates the coding process more naturally, and we found that it provided a 0.4% improvement in performance. Detailed performance metrics are

provided in Appendix A.

Training Setting	Average Score
Python	76.4(-1.2)
C++	76.6(-1.0)
Java	76.4(-1.2)
Python <sub>3×</sub>	76.1(-1.5)
MPL <sub>sampled</sub>	77.1(-0.5)
MPL	<b>77.6</b>
w/o. Virtual Running	77.2(-0.4)

Table 4: Ablation studies on various components in MPL, we use LLaMA3-8B and function-prompt as default setting here.

## 5 Analysis

### 5.1 Which Backbone Is More Suitable For Code-style Input?

Intuitively, IE tasks are knowledge-intensive and require not only strong natural language (NL) understanding to capture semantic meaning but also robust programming language (PL) capabilities to generate structured knowledge. In this subsection, we evaluate which LLM backbone is most suited for code-style inputs in the context of IE tasks. The tested LLMs include StarCoder-v2, LLaMa2-7B, and LLaMA3-8B. For a deeper analysis, we roughly measure the language capabilities based on the performance in the MMLU (Hendrycks et al., 2021) (measuring NL ability) and HumanEval (Chen et al., 2021a) (measuring PL ability)<sup>6</sup>.

From Table 2 we can see that: 1) StarCoder-v2 performs the worst in IE tasks due to its limited natural language understanding capabilities, despite excelling in PL tasks. On the other hand, LLaMa2-7B achieves significantly better results, even with weaker PL capabilities. 2) LLaMA3-8B achieves the best result, owing to its superior NL and PL understanding, making it the ideal backbone for code-style inputs. In conclusion, the NL ability is more important than PL ability in the case of IE with code-style inputs. However, proper PL understanding ability could further enhance the overall performance.

<sup>6</sup>The results of MMLU and HumanEval are cited from StarCoder-V2 Tech Report, LLaAa2 Tech Report and LLaAa3 Tech Report.

### 5.2 Disentangling Programming Language Diversity from Ensemble Effects

To rigorously assess whether the performance gains of our MPL framework stem from true programming language diversity or from general ensemble effects, we designed a series of controlled baseline variants. These include: (1) training with a single programming language (Python-only); (2) ensembling multiple models trained with different random seeds using the same language; (3) prompt augmentation by varying label order or format while tripling data volume; and (4) ensembling independently trained models using different programming languages (Python, C++, Java). These variants isolate different sources of potential performance improvements, including random initialization effects, superficial prompt perturbations, and structural diversity introduced through programming language variation. Our unified MPL design incorporates multi-language supervision during training, aiming to embed more diverse code-level semantics into the model’s representation.

Model Variant	Avg. F1
Single PL (Python only)	76.40
Seed Ensemble (Python × 3 seeds)	76.60
Label Reordering + 3× data	76.22
Prompt Format Variation + 3× data	76.63
PL Ensemble (Python, C++, Java)	76.73
MPL (ours)	<b>77.60</b>

Table 5: Performance comparison of MPL with various ablation baselines.

As summarized in Table 5, ensembling models trained with different random seeds (76.60 F1) or applying prompt variations with expanded data (76.22–76.63 F1) yields only marginal gains over the single-language baseline (76.40 F1). In contrast, ensembling across different programming languages improves performance more noticeably (76.73 F1), suggesting that language diversity contributes beyond initialization noise or prompt form. Notably, our unified MPL training outperforms all variants (77.60 F1), achieving a +1.2 improvement over the Python-only baseline. These results confirm that programming language diversity introduces deeper structural and syntactic variability, fostering more robust task understanding and generalization. The observed advantage cannot be fully explained by general ensembling or superficial vari-

ations, validating the core design principle of MPL.

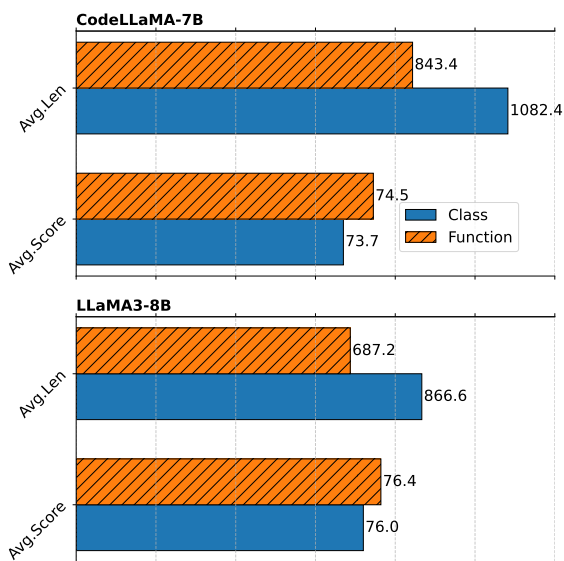


Figure 3: Performance and training statistics for different LLMs with various input formats. **Avg. Score** represents the average score across all datasets, while **Avg. Len** denotes the average input length after tokenization with the corresponding LLM’s tokenizer. Notably, the same prompt yields different **Avg. Len** values across models due to variations in their tokenizer configurations. Detailed counts are provided in Appendix D

### 5.3 Function-prompt vs. Class-prompt

Another contribution of this research is the introduction of function-prompt as the alternative to class-prompt. In this subsection, we provided an in-depth analysis comparing the performance and training efficiency of these two prompts. Specifically, we report the average score and the average input length among all training datasets. Figure 3 presents the experimental results using different input formats. In terms of performance, our function-prompt consistently outperforms the class-prompt across various LLMs. Regarding training efficiency, the average input length of function-prompt is reduced around 20% compared to class-prompt, highlighting its efficiency in processing inputs.

### 5.4 Complementarity Between Different Programming Languages

This subsection presents detailed performance analysis of different PLs and aggregation methods within our MPL framework. We investigate single PL predictions and two aggregation methods: **Voting** (Default), which combines predictions from all three PLs to determine the final output, and **Union**,

where a prediction is considered correct if any of the PLs produce the correct result, representing the union of all predictions. The results are shown in Table 6 and we can see that: 1) The performance benefit of **Voting** over individual PL predictions is minimal. This suggests that MPL effectively captures and integrates the diverse semantic meanings and formats inherent in different PLs, leading to robust performance. Consequently, selecting a single PL during testing can yield desirable results with reduced inference costs. 2) Aggregating results via **Union** significantly improves average scores, confirming our hypothesis that different PLs provide complementary strengths for varied IE tasks and datasets. Although **Voting** has been the default, these findings suggest exploring more effective aggregation methods could be beneficial.

Model	Prediction Setting	Avg.Score
MPL (CodeLLaMA)	Python	75.6
	C++	75.8
	Java	75.4
	<b>Voting</b>	76.1
	<b>Union</b>	<b>78.4(+2.3)</b>
MPL (LLaMa3)	Python	77.2
	C++	77.3
	Java	77.3
	<b>Voting</b>	77.6
	<b>Union</b>	<b>79.5(+1.9)</b>

Table 6: The average score with different prediction settings. The results are all obtained from well-trained MPL-8B models.

Besides, we compute the jaccard similarity coefficient (Jaccard, 1912) to evaluate the similarities between different PLs’ predictions. First, we compute the Jaccard similarity between each pair of PLs, and then we show the average coefficients in Table 7. We also analysis the performance gap between **Voting** and **Union** for a more comprehensive evaluation. From the results we can observe that: 1) For simple tasks like NER, the models perform well and the predictions across PLs are quite similar, showing minimal differences between ensemble methods. 2) For more complex tasks such as RE and EE, we not only observe a significant drop in overall model performance but also a substantial discrepancy in prediction similarity. This indicates that different languages excel at handling different tasks and datasets, leading to lower prediction simi-



larity and more substantial performance differences between different ensemble methods. This finding strongly supports our motivation that different PLs have distinct characteristics and excel in different tasks. Through the design of MPL, we effectively harness the potential for cross-linguistic enhancement, thereby improving overall performance.

	Jaccard Similarity	MPL	Union-Voting
<b>ACE05-NER</b>	0.98	91.4	2.25
<b>BC5CDR</b>	0.99	89.6	1.37
<b>CONLL03</b>	0.99	93.5	1.45
<b>DIANN</b>	0.98	85.4	0.08
<b>ACE05-RE</b>	0.93	70.8	16.25
<b>ACE05-EAE</b>	0.94	72.8	2.93
<b>RAMS</b>	0.88	50.9	3.55
<b>ACE05-EE</b>	0.97	72.7	3.67

Table 7: We calculate the jaccard similarity between each pair of PLs and report the average coefficients, the higher, the similar. We also show the performance gap between voting and union.

## 6 Related Work

Information Extraction involves entity recognition, relation extraction, event extraction and other related tasks (Li et al., 2024a; Vajjala and Balasubramaniam, 2022; Li et al., 2023b,c; Gao et al., 2024; Lu et al., 2021; Wang et al., 2022a). There are plenty works focusing on information extraction in the era of large language model. Earlier research mainly evaluated the performances on various large language models under few-shot and zero-shot settings. For example, Li et al. (2023a); Han et al. (2023); Li et al. (2024b) evaluated LLMs’ performance and other dimensions to provide systemically analysis, while Wadhwa et al. (2023); Xu et al. (2023); Zhu et al. (2023); Jiang et al. (2024) discussed the opportunities and challenges for IE field research in the era of large language model.

Recent studies have explored the use of code-style inputs (Chen et al., 2021b; Roziere et al., 2023; Li et al., 2023e; Zheng et al., 2023) for IE tasks, aiming to structure these tasks more formally and facilitate LLMs in generating structured outputs more accurately. CodeIE(Li et al., 2023d; Wang et al., 2022b) introduced code-style representations for various IE tasks under few-shot settings. Code4UIE(Guo et al., 2023) proposed a

retrieval-augmented code generation framework, integrating class-prompting and multiple example retrieval strategies to enhance few-shot performance. CodeKCG (Bi et al., 2024) reformulated natural language into code-style formats for generative knowledge graph construction, incorporating rationale-enhanced generation as an intermediate step to provide additional contextual information for unseen examples. Additionally, some studies have explored fine-tuning LLMs with code-style inputs. GoLLIE(Sainz et al., 2023) introduced fine-grained label descriptions and candidate selection to improve supervised and zero-shot performance. KnownCoder(Li et al., 2024c) first designed a schema understanding phase with millions of pre-training instances to enhance LLM comprehension of code-style inputs, followed by a supervised fine-tuning phase on 33 domain-specific datasets, referred to as the schema-following phase.

## 7 Conclusion

In this paper, we propose MPL, a novel framework that leverages multiple programming languages with large language models for information extraction. We first propose function-prompt with virtual running component, a more lightweight and effective code-style simulate method compared to class-prompt. Then, MPL reformulates a given IE task and textual input into a code-style input with three programming languages, i.e., Python, C++ and Java to explore the potential of cooperation between them. Extensive experimental results on various IE datasets not only validate the effectiveness of MPL but also provide a thorough analysis to help readers better understand our method. To facilitate future research, we will release our code and necessary files later.

### Limitation

Although our research explores the potential of using multiple programming languages and achieves significantly better performance than previous works, training efficiency remains a notable drawback, leading to more than twice the training cost compared to the single-PL setting. Future work will focus on reducing training time under the MPL framework while maintaining performance.

### Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.

62276089), the Natural Science Foundation of Tianjin (Grant No. 24JCQJC00200), the Natural Science Foundation of Tianjin (Grant No. 24JC-QNJC01230), the Natural Science Foundation of Hebei Province (Grant No. F2024202064), the Science Research Project of Hebei Education Department (Grant No. BJ2025004), the Ministry of Human Resources and Social Security of China (Grant No. RSTH-2023-135-1), and the Science and Technology Program of Hebei Province (Grant No. 24464401D).

## References

- Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–16.
- Bruno Blanchet. 1999. Escape analysis for object-oriented languages: application to java. *Acm Sigplan Notices*, 34(10):20–34.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Matusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021a. [Evaluating large language models trained on code](#). *CoRR*, abs/2107.03374.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021b. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, pages 1169–1179.
- Leon Derczynski, Eric Nichols, Marieke Van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.
- Rezarta Islamaj Dogan and Zhiyong Lu. 2012. An improved corpus of disease mentions in pubmed citations. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 91–99.
- J Michael Dunn. 2019. Natural language versus formal language. *New Essays on Belnap-Dunn Logic*, pages 13–19.
- Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2019. Multi-sentence argument linking. *arXiv preprint arXiv:1911.03766*.
- Hermenegildo Fabregat, Juan Martinez-Romo, and Lourdes Araujo. 2018. Overview of the diann task: Disability annotation task. In *IberEval@ SEPLN*, pages 1–14.
- Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2024. Promptre: Weakly-supervised document-level relation extraction via prompting-based data programming. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 132–145.
- Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, et al. 2023. Retrieval-augmented code generation for universal information extraction. *arXiv preprint arXiv:2311.02962*.
- Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance,

- evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Gustaf Holst and Felix Dobsław. 2021. On the importance and shortcomings of code readability metrics: A case study on reactive programming. *arXiv preprint arXiv:2110.15246*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.
- Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. *arXiv preprint arXiv:2402.10744*.
- Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt’s information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.
- Bo Li, Wei Ye, Quansen Wang, Wen Zhao, and Shikun Zhang. 2024a. Labels need prompts too: mask matching for natural language understanding tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18453–18461.
- Bo Li, Wei Ye, Jinglei Zhang, and Shikun Zhang. 2023b. Reviewing labels: Label graph network with top-k prediction set for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13051–13058.
- Bo Li, Dingyao Yu, Wei Ye, Jinglei Zhang, and Shikun Zhang. 2023c. Sequence generation with label augmentation for relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13043–13050.
- Guozheng Li, Zijie Xu, Ziyu Shang, Jiajun Liu, Ke Ji, and Yikai Guo. 2024b. Empirical analysis of dialogue relation extraction with large language models. *arXiv preprint arXiv:2404.17802*.
- Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023d. Codeie: Large code generation models are better few-shot information extractors. *arXiv preprint arXiv:2305.05711*.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023e. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.
- Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *arXiv preprint arXiv:2104.05919*.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023f. Self-alignment with instruction back-translation. *arXiv preprint arXiv:2308.06259*.
- Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xi-ang Li, Zhilei Hu, et al. 2024c. Knowcoder: Coding structured knowledge into llms for universal information extraction. *arXiv preprint arXiv:2403.07969*.
- Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.
- Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806.
- Navapat Nananukul and Wichayaporn Wongkamjan. 2024. What if red can talk? dynamic dialogue generation using large language models. *arXiv preprint arXiv:2407.20382*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.



- Danny CC Poo, Derek Kiong, and Swarnalatha Ashok. 2008. *Object-oriented programming and Java*. Springer Science & Business Media.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.
- Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Dipanjan Sarker. 2016. *Text analytics with python*, volume 2. Springer.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhunoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deep-speed and megatron to train megatron-turing nl-g 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Bjarne Stroustrup. 2013. *The C++ programming language*. Pearson Education.
- Bjarne Stroustrup. 2014. *Programming: principles and practice using C++*. Pearson Education.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Sowmya Vajjala and Ramya Balasubramaniam. 2022. What do we really know about state of the art ner? *arXiv preprint arXiv:2205.00034*.
- Peter Van Roy and Seif Haridi. 2004. *Concepts, techniques, and models of computer programming*. MIT press.
- Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.
- Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022a. Query and extract: Refining event extraction as type-oriented binary decoding. In *ACL (Findings)*.
- Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot event structure prediction. *arXiv preprint arXiv:2210.12810*.
- Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wieggers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016:baw032.
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. [Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19368–19376. AAAI Press.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.
- Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.
- Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *arXiv preprint arXiv:2305.13168*.