



ARTICLE

TIPS: Tailored Information Extraction in Public Security Using Domain-Enhanced Large Language Model

Yue Liu¹, Qinglang Guo², Chunyao Yang¹ and Yong Liao^{1,*}

¹School of Cyber Science and Technology, University of Science and Technology of China, Hefei, 230026, China

²National Engineering Research Center for Public Safety Risk Perception and Control by Big Data, China Academy of Electronics and Information Technology, Beijing, 100041, China

*Corresponding Author: Yong Liao. Email: ylliao@ustc.edu.cn

Received: 29 October 2024; Accepted: 17 February 2025; Published: 16 April 2025

ABSTRACT: Processing police incident data in public security involves complex natural language processing (NLP) tasks, including information extraction. This data contains extensive entity information—such as people, locations, and events—while also involving reasoning tasks like personnel classification, relationship judgment, and implicit inference. Moreover, utilizing models for extracting information from police incident data poses a significant challenge—data scarcity, which limits the effectiveness of traditional rule-based and machine-learning methods. To address these, we propose TIPS. In collaboration with public security experts, we used de-identified police incident data to create templates that enable large language models (LLMs) to populate data slots and generate simulated data, enhancing data density and diversity. We then designed schemas to efficiently manage complex extraction and reasoning tasks, constructing a high-quality dataset and fine-tuning multiple open-source LLMs. Experiments showed that the fine-tuned ChatGLM-4-9B model achieved an F1 score of 87.14%, nearly 30% higher than the base model, significantly reducing error rates. Manual corrections further improved performance by 9.39%. This study demonstrates that combining large-scale pre-trained models with limited high-quality domain-specific data can greatly enhance information extraction in low-resource environments, offering a new approach for intelligent public security applications.

KEYWORDS: Public security; information extraction; large language model; prompt engineering

1 Introduction

The rapid advancement of information technology, marked by faster information dissemination, exponential data growth, and increasingly diverse forms of crime, has introduced intricate challenges to public security governance. Big data and artificial intelligence have emerged as indispensable tools for tackling the complexities of massive policing data, offering robust support for the development of intelligent policing systems. In public security operations, precise management of policing data is crucial for achieving efficient performance. Currently, the diversification and accessibility of reporting channels allow police to input substantial volumes of incident data into internal systems on a daily basis. While this improves the comprehensiveness of incident data collection, it also results in rapid data expansion, unstructured content, and inconsistent quality. These challenges complicate manual filtering and information extraction, potentially delaying analysis or leading to omissions, which in turn impact the efficiency of police operations and the accuracy of decision-making. These incident reports primarily consist of unstructured natural language text, encompassing key information such as individuals, events, objects, time, locations, and their interrelated connections. Efficiently and accurately extracting key information from incident data serves as



a solid foundation for decision support in subsequent tasks, including incident analysis, risk assessment, and command and dispatch. Therefore, Leveraging information technology to extract key information from police incident data and facilitate efficient comprehension of large-scale text is essential for strengthening policing capabilities.

Information extraction refers to extracting valuable information from data of various sources and structures [1]. Its core objective is to transform unstructured or semi-structured data into structured data for more effective storage, management, and utilization [2,3]. Traditional small-scale deep learning models for information extraction perform well after learning features from large-scale, high-quality data [4–6]. However, their performance often declines in specific domains with scarce data resources [7,8]. In public security, the sensitivity and confidentiality of business data require strict protection regulations, limiting real data use to internal police networks. This poses challenges for obtaining specialized data. And even when relevant data is obtained, the high data requirements of traditional models compel researchers to perform extensive fieldwork, such as manual preprocessing and annotation, which is both time-consuming and labor-intensive. As a result, publicly available annotated data in the public security domain is highly limited. This makes traditional information extraction models, which rely on large volumes of high-quality data, inefficient for information extraction tasks in low-resource environments. For example, these models often struggle to fully capture domain-specific knowledge and are prone to overfitting on limited datasets [9].

Moreover, traditional information extraction models primarily focus on a single task, such as named entity recognition (NER) [10,11], relation extraction (RE) [12,13], or event extraction (EE) [14,15]. While some studies have combined two tasks [16,17], they face challenges such as increased model complexity and higher training costs. The complexity arises from the intricate interactions between tasks—e.g., event triggers may depend on recognized entities, while relation extraction relies on connections between entities—making research on integrating all three tasks both limited and difficult to generalize. In the public security domain, data contains rich and critical case information, including entities (e.g., people and objects), relationships (e.g., person-object and person-person), and event elements (e.g., time, location, and human-event relationships). Existing single-task and joint extraction models fail to fully and efficiently meet the demands of comprehensive information extraction in this context. In conclusion, traditional information extraction models encounter significant challenges in processing data within the public security domain: **(1) Limited ability to learn domain-specific knowledge under low-resource conditions.** Public security data often contains highly fine-grained information, and training on small datasets frequently results in overfitting. **(2) Inefficiency in addressing the information-intensive demands.** On one hand, public security data is inherently dense, requiring the simultaneous extraction of diverse entities, relationships, and events, which is rarely achieved by traditional models. On the other hand, existing joint extraction models involve high training costs and are difficult to generalize, further hindering their practical application. Therefore, in low-resource and restricted data environments, employing traditional models for joint information extraction in the public security domain presents significant challenges.

To address the challenges faced by traditional information extraction models, this study introduces a template-based synthetic dataset and employs a method that combines domain-specific extraction frameworks with LoRA [18] technology to fine-tune large language models for information extraction in the public security domain:

1. **Leveraging LLMs for Low-Resource Extraction:** To address the limitations of traditional neural networks in low-resource scenarios, this study utilizes LLMs fine-tuned with advanced techniques for information extraction in the public security domain. With their robust language comprehension capabilities, LLMs can effectively adapt to task-specific requirements using minimal domain-specific data, while exhibiting strong generalization on unseen data.

2. **Construction of a Realistic Joint Extraction Dataset:** To mitigate the scarcity of sharable data resources in the public security domain, this study generates simulated data grounded in realistic incident scenarios and constructs a joint extraction dataset through human-machine collaboration. The dataset incorporates tasks such as named entity recognition, relation extraction, and event extraction, facilitating LoRA fine-tuning of the model. By leveraging the joint dataset alongside the reasoning capabilities of large language models, this approach enables multi-task joint extraction of police information, efficiently producing domain-specific extraction models with minimal data and computational resources.
3. **Domain-Enhanced Controlled Information Extraction:** To address the complex requirements of information extraction in police operations, this paper proposes a domain-specific extraction schema based on prompt engineering. The schema includes entities, attributes, and relationships commonly found in public security incidents, such as the reporter's name and ID card number. It allows for the addition or removal of extraction targets via prompts, enabling the model to transfer its extraction capabilities to similar data and ensuring scalability. By incorporating the schema into input prompts, the approach enhances the model's comprehension of professional texts and enables controlled generative information extraction.
4. **Application in Practical Police Operations:** The joint dataset construction method and extraction framework proposed in this paper have been implemented in the big data system of a specific city's public security department. They enable the extraction of entities, attributes, and relationships from diverse data sources such as police incident records and public opinion information. The extracted results are used for data analysis and the construction of knowledge graphs related to public security incidents, offering practical support to frontline police officers in evaluating and analyzing such incidents.

2 Related Work

With the continuous development of the concept of intelligent public security, information extraction [19], as a prerequisite for understanding and analyzing public security incidents, has gradually gained attention. Many studies have developed methods for public security information extraction based on artificial intelligence technology [20–22]. As early as 2008, Ku et al. [23] developed an information extraction system by combining the principles of natural language processing and cognitive interviewing, aiming at extracting crime-related information from police and eyewitness narratives. In recent years, with the development of language models, research on information extraction based on the BERT model has become mainstream. Deng et al. [24] proposed a BERT-based [25] GRU-CRF [26,27] model for entity extraction in drug-related and money transfer incidents, achieving SOTA compared to DMCNN [28]; Cui et al. [29] designed a BERT-BiGRU-SelfAtt-CRF model for police entity recognition, achieving optimal F1 on a public dataset; Zhao et al. [30] utilized BERT to encode semantic information and target entities, applied a GCN to model dependency graphs of relationship labels, and successfully extracted relationships in police events.

Traditional AI models (such as BERT and CRF) still face challenges in the application of police data processing. For instance, these methods have poor scalability when dealing with large-scale data, rely heavily on a large amount of annotated data, and perform inadequately when handling highly specialized domain-specific language. With advancements in LLMs, large-scale models like GPT3.5 [31] have demonstrated exceptional language understanding and generation capabilities, excelling in downstream tasks such as machine translation, reading comprehension, and intelligent customer service [32–34]. At the same time, recent studies have proposed LLM-based methods for information extraction [35], which can achieve good extraction results without the need for full-volume parameter training or only through prompt word engineering [36–39].

Therefore, studying LLM-based information extraction methods in public security is crucial, as their advancements can greatly enhance extraction accuracy and efficiency. This study aims to develop an efficient method for applying LLMs to public security, overcoming the limitations of traditional models and providing intelligent tools to improve police response speed and decision-making quality. Through in-depth research, this method is expected to significantly improve information extraction and lay a technical foundation for intelligent public security.

3 Methodology

For information extraction in the public security field, this paper proposes a joint information extraction method based on LLMs. By leveraging the advanced text comprehension and generation capabilities of LLM, combined with model fine-tuning techniques, a domain-specific LLM is developed to extract various types of key information from police incident data. As shown in Fig. 1, the process begins with strictly anonymized real police data, used in a human-machine collaborative manner to generate simulated incident data for building a joint extraction dataset. A computationally efficient and deployable model is then selected and fine-tuned using this dataset to better align with the public security domain. Prompt engineering is applied to activate the model's extraction capabilities, with domain-specific prompts and templates constraining its generative output. Finally, unstructured police incident data is input into the trained model, which outputs structured key information.

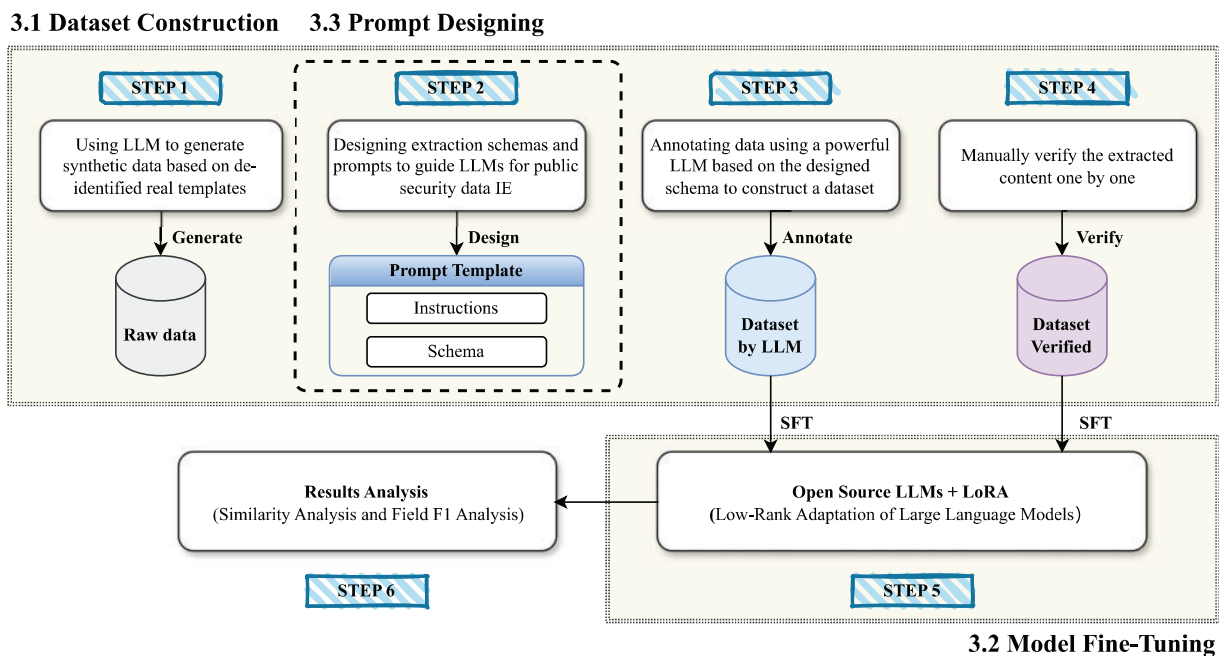


Figure 1: Framework of TIPS. It consists of three main modules: Dataset construction, model selection and fine-tuning, and prompt designing. The entire process includes six steps as shown in the figure, where Step 2 constitutes the prompt designing module. The same prompt templates are used during both dataset construction and model inference

3.1 Dataset Construction

In this paper, we constructed a police text extraction dataset comprising 1000 entries across various case categories, including marriage and family, neighborhood disputes, consumption, property, land, labor

disputes, traffic accidents, intellectual property, contracts, telecommunication fraud, theft, and more. The dataset construction process is illustrated in Fig. 2.

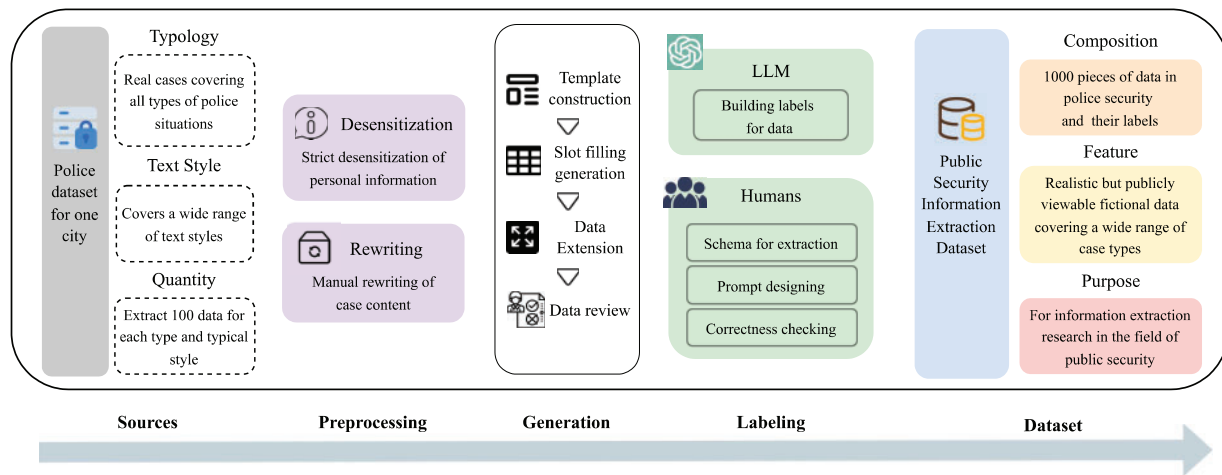


Figure 2: Process of dataset construction. One hundred real police data from a specific city (Sources) undergo desensitization and manual rewriting during the preprocessing. Through template construction, slot filling, and data extension, 1000 synthetic yet realistically styled police cases are generated (Generation). The relevant data is then labeled via collaboration between large language models and human annotators, involving the design of schema and prompts, as well as the correctness verification of automated labeling (Labeling). The final dataset consists of 1000 high-quality synthetic police incident records for information extraction, supporting research in information extraction within the public security domain (Dataset)

To ensure data validity while adhering to privacy protection requirements and data export regulations, we conducted an on-site investigation of a municipal police incident database, as shown in the Sources section. Guided by police officers, we analyzed the structure and style of incident records and carefully selected 100 real police incident cases from the database. The selection was based on two main criteria: (1) Case Types: The accessible data was limited to civil cases. Following guidance from legal experts, we referred to 17 categories of civil disputes and selected representative cases for each category. Cases that did not fit these categories were classified as “Other”. (2) Text Characteristics: Incident records varied in length, information density, and content. While personal privacy information was consistently recorded, other details depended on the complexity of the incident and the reporter’s ability to provide information. Additionally, slight stylistic differences arose due to variations in individual recorders’ styles. Considering these factors, we selected cases representing variations in text length, information density, and recording style to ensure a diverse and comprehensive dataset. Subsequently, we anonymized personal information in the selected records and manually rewrote event details to ensure significant differences from the original cases, guaranteeing that the exported data did not represent real incidents. These 100 records were then used as templates, with key fields blanked out and target slots designed. The templates were input into a closed-source LLM, guided to fill in the blanks and generate realistic but fictitious police incident data. Each generated record was further used as input to create ten additional records with similar styles, resulting in 1000 valid and exportable records. Then all records were manually reviewed for logical consistency, and any invalid data was carefully corrected.

To address concerns about dataset diversity and complexity, we acknowledge that the synthetic dataset may not fully represent real-world police reports. Due to the sensitive nature of police data, obtaining large amounts of desensitized data requires lengthy approval processes, which could not be secured within

the current timeframe. Future work will focus on expanding the dataset by incorporating records from diverse regions and incident types through collaboration with multiple police departments and advanced data augmentation techniques, ensuring a more comprehensive representation of real-world scenarios.

After constructing the valid incident dataset, we invited frontline police officers and police academy cadets to confirm the extraction requirements and develop an information extraction schema. This schema includes fields for the names, ID numbers, contact information, addresses, and birthplaces of the reporter, disputants, and responding officers, as well as related case information such as dispute parties and the time of the report. The extraction process encompasses not only entity extraction (e.g., names, ID numbers, contact information) but also relation extraction (e.g., associations between reporters and disputants, or between ID numbers and reporters) and event extraction (e.g., identifying reporters, case occurrence times, locations, and related event elements). Consequently, the dataset generated based on this schema is designed to train models capable of performing joint extraction of entities, relations, and events. To reduce labor and time costs, the data annotation process adopts a human-machine collaborative approach. First, prompt words for the extraction schema (outlined in [Section 3.3](#)) were designed to guide the closed-source LLMs in labeling 1000 raw data entries. The output dataset was then manually reviewed to verify the accuracy of the labeling and correct any errors. This approach effectively reduces manual effort while ensuring the accuracy of the annotations. Examples of the generated dataset are shown in [Fig. A1](#).

3.2 Model Fine-Tuning

In the current wave of artificial intelligence, various types of LLMs are competing. Since the introduction and popularity of ChatGPT, similar models have emerged, demonstrating superior performance across domains. Despite the impressive performance of LLMs, there remains significant room for improvement in specific industries and tasks. To achieve optimal performance, fine-tuning is essential for task-specific adaptation. This process not only enables deeper domain understanding but also improves resource efficiency to meet application demands. Unlike traditional comprehensive fine-tuning, which is resource-intensive due to LLMs' large parameter counts, full fine-tuning requires substantial computational power, as well as significant time and cost. Therefore, under constrained resources, adopting efficient fine-tuning strategies is crucial for specific downstream tasks.

In this study, we employ LoRA technology to fine-tune the selected base model for specialized information extraction tasks in the public security domain. LoRA fine-tuning technology works by freezing the pre-trained model's weights and introducing two matrices to replace the parameters. Thus, during fine-tuning, only these two matrices are updated. By leveraging low-rank decomposition to simulate parameter changes, LoRA enables efficient training with minimal parameter updates.

3.3 Prompt Designing

In both [Sections 3.1](#) and [3.3](#) of this paper, a prompt template designed for information extraction in the public security domain is utilized. This prompt template guides information extraction as a generative task, constraining the model's output to ensure controllability and structured extraction of key information from unstructured police incident text. To meet these requirements, the prompt template is tailored to the practical needs of public security scenarios, providing a framework for extracting information and enhancing the accuracy of domain-specific extraction by the fine-tuned LLM.

The prompt template contains five components: Roles, Skills, Functions, Schema, and Police incident text. The Role module instructs the model to act as an extractor specialized in public security information. The Skills module highlights the key capabilities required for the role, activating them to effectively process relevant data. The Function module defines additional requirements, such as output format and exception

handling, ensuring accurate content generation. The Schema module provides a predefined framework for extracting information so the model can adhere to user-specified requirements. Finally, the Text module contains the input text for extraction. The design framework for the prompt template is illustrated in Fig. 3, showcasing its role in guiding and constraining generative extraction.

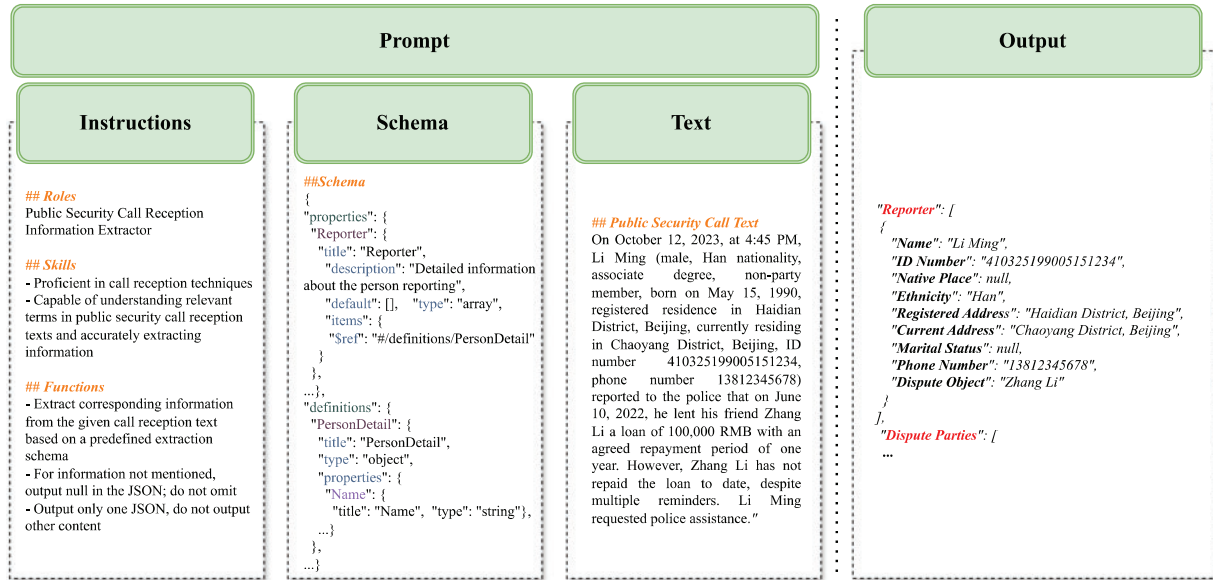


Figure 3: Prompt Template Design Framework. The framework for the prompt template, as shown in the figure, includes three parts: Instructions, Schema, and Text. The prompt template is input into the fine-tuned LLM, which then outputs the extraction results

4 Experiments

4.1 Baseline and Metrics

Considering our task is centered on natural language understanding and structured text generation, we tend to choose the currently known powerful models for testing. Additionally, since our task requires JSON format output, models with better instruction-following abilities are anticipated to yield superior results. Thus, we have selected the following models for our experiment:

1. Qwen2-7B-Instruct [40] is the new series of Qwen large language models.
2. GLM-4-9B-Chat [41] is the open-source version of the latest generation of pre-trained models in the GLM-4 series launched by Zhipu AI, claimed to have shown superior performance beyond LLaMA-3-8B.
3. Meta-Llama-3-8B-Instruct [42] is the next iteration of the open-access Llama family, with a context length of 8K tokens.
4. Mistral-7B-Instruct-v0.3 [43] is an instruct fine-tuned version of the Mistral-7B-v0.3.

We also utilize various methods to evaluate the performance of the fine-tuned models. Initially, we apply standard similarity metrics used in NLP tasks, such as BLEU [44] and ROUGE [45], to qualitatively assess the extraction results, offering a preliminary evaluation of the overall performance. Subsequently, we conduct quantitative analysis on fields, calculating metrics like F1 [46] to provide a comprehensive assessment of the fine-tuned model's capabilities. During quantitative testing, the model's output may contain various errors, which we addressed as follows:

1. The model, in addition to outputting JSON, also produced some textual content. We considered this as not fully adhering to the given instructions. However, if the text contains extractable and processable information, it is still acceptable. Therefore, we attempt to extract the required JSON text through format matching. If extraction fails, we generate fake JSON data filled with the <FAKE> identifier. Regardless of the success of the extraction, it will be recorded as an Extra Text error.
2. The model outputted the required JSON but missed some keys (*Missing Keys*) or added extra keys (*Extra Keys*). We filled in the missing keys with <PADDING> identifier and removed the additional keys added by the model.
3. The model outputted JSON, but some keys had incorrect types (*Wrong Type*). We corrected these and filled the gaps with <PADDING> identifier.

These errors are also documented in evaluating the model's performance.

4.2 Experiments Setup

All experiments are conducted on Ubuntu 20.04 with 2 Intel Xeon Gold 5520 CPUs, and 8 NVIDIA Tesla V100 32 GB GPUs, using LLaMA-Factory [47] as the primary fine-tuning framework.

In the process of Lora fine-tuning, we utilize AdamW [48] as the optimizer, with a learning rate of 5e-5, 5 epochs, and a batch size of 2. In the evaluation phase, we set the temperature to 0.3–0.8 and top-p to 0.6–0.8 to assess the model's output under various conditions.

In the process of constructing our dataset, we initially utilized a powerful LLM to automatically annotate the data, resulting in what we refer to as Generated Dataset. This dataset was subsequently divided into training and testing subsets at a ratio of 0.85:0.15 to facilitate model training and evaluation. Following the automatic annotation, we meticulously conducted a manual verification of the annotations generated by the LLM, reviewing them one by one to ensure accuracy and correctness. This rigorous process led to the creation of a Verified Dataset, which reflects a higher degree of accuracy and reliability. During fine-tuning, we used 10% of the training set as the validation set to ensure better fine-tuning results. All evaluations are conducted on the verified test set.

4.3 Results Analysis

We first conducted an in-depth analysis of the temperature and top-p parameters using the Qwen2 model, which was chosen as a representative model for our experiments to determine the optimal settings for maximizing the quality of generated outputs. As illustrated by the data presented in Table 1, we observed that the model achieved near-optimal performance when the temperature parameter was set to 0.5 and the top-p parameter was configured at 0.8. These settings closely align with the objectives of our specific tasks.

In tasks where LLMs are employed to extract information and produce structured outputs, it is crucial to ensure that the generated content is not only stable and consistent but also retains a sufficient level of diversity. The combination of a temperature of 0.5 and a top-p of 0.8 successfully strikes a balance between these requirements, leading to improved overall performance in our tests. Consequently, in all subsequent comparative analyses, we have consistently utilized these parameters to maintain the desired quality of generation.

Table 2 presents a detailed comparison of the results obtained from the four models both before and after the fine-tuning process, with the best-performing results highlighted in bold for clarity.

Table 1: Performance metrics for different temperature and Top-p values. Bold values indicate the highest scores under specific parameter combinations

Top-p	Temperature	Metrics		
		BLEU-4	ROUGE-1	F1
0.6	0.3	93.83	95.77	81.81
0.7		94.71	96.32	83.31
0.8	0.3	94.83	96.44	83.79
	0.4	94.22	96.13	84.13
	0.5	94.75	96.34	84.75
	0.6	91.85	95.02	81.50
	0.7	93.16	95.40	82.23
	0.8	92.85	95.33	82.71

Table 2: Performance metrics and errors for different models. Bold values denote the best performance of base models and SFT models

Models		Metrics			Errors			
		BLEU-4	ROUGE-1	F1	Extra text	Missing keys	Extra keys	Wrong type
Base	Mistral	58.34	67.60	47.33	106	765	183	0
	LLaMA3	60.77	68.17	61.46	144	1001	68	4
	Qwen2	51.83	69.13	41.06	0	2271	137	15
	GLM-4	47.33	71.11	60.17	0	1529	75	17
SFT	Mistral	91.32	94.68	82.32	0	1	1	0
	LLaMA3	92.84	95.70	84.58	0	0	0	0
	Qwen2	94.75	96.34	84.75	0	0	0	0
	GLM-4	94.61	96.62	87.14	0	0	0	0

Before fine-tuning, the outputs generated by LLaMA3 for all entries in the test set were not limited to the expected JSON format; instead, they included additional, unintended text elements. A similar issue was observed with the Mistral model, although this problem was somewhat less pronounced in comparison to LLaMA3. In contrast, both Qwen2 and GLM-4 models consistently succeeded in generating outputs strictly in the JSON format, without any extraneous text. We believe this superior performance can be attributed to the stronger support for the Chinese language provided by Qwen2 and GLM-4, which stems from their significantly larger and more comprehensive Chinese-language training datasets. These datasets enhance the models' understanding and compliance with Chinese-language instructions.

However, it is also worth noting that, possibly due to the inherent limitations of the models, Qwen2 and GLM-4 demonstrated slightly lower performance compared to LLaMA3 and Mistral when it came to the specialized task of extracting information from police report data. Among these two, GLM-4 outperformed Qwen2, a result that can be attributed to its larger parameter size.

After fine-tuning the models with domain-specific data, we observed a notable change characterized by a significant reduction in the number of errors across all models. Despite this improvement, the Mistral

model still exhibited a minor error. Upon closer inspection and verification, it was identified that one of Mistral's outputs had "ID number" replaced with "ID card number." These two terms are nearly identical when translated into Chinese, suggesting that Mistral's proficiency and understanding of nuanced language differences in Chinese require further enhancement.

Further analysis of the specific evaluation metrics revealed that the Qwen2 model's F1 score experienced a remarkable increase of approximately 43%, demonstrating substantial gains in performance. Meanwhile, the other three models also showed notable improvements, with increases ranging from 25% to 35%. Among the models tested, GLM-4, which has the largest parameter size, achieved the highest performance level, with Qwen2 coming in as a close second. This performance hierarchy indicates that even minimal adaptation using domain-specific data can substantially enhance the models' abilities in information extraction tasks. These findings indirectly reflect the high quality and relevance of the dataset used for fine-tuning, as it effectively contributed to the models' improved performance.

Moreover, we calculated the F1 score for each field individually to conduct a more detailed investigation into the specific areas where the models underperformed. The results of this analysis are presented in Fig. 4. Among the fields analyzed, the "Job Title" field exhibited the lowest performance, with an F1 score of approximately 0.4. We hypothesize that this low score is primarily due to the limited presence of job title information in the training set, as most entries did not explicitly mention the dispatcher's job title.

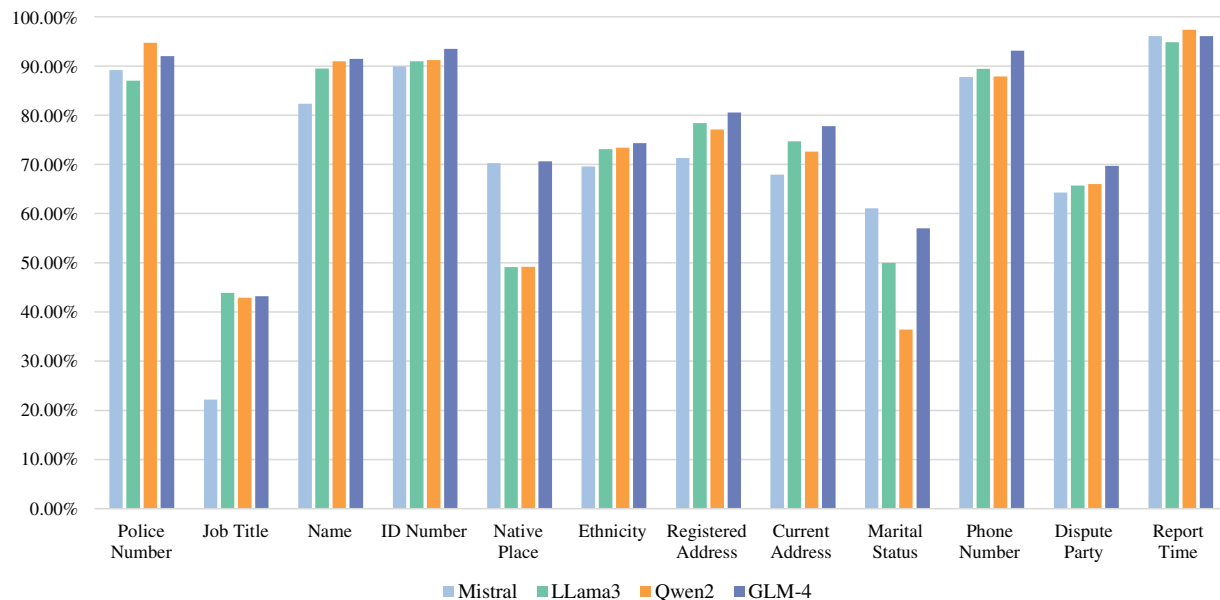


Figure 4: F1 scores by fields for different models

Similarly, as demonstrated in Table 3, fields such as "Ethnicity," "Native Place," "Police Number," and "Marital Status" are relatively scarce in the dataset, much like the "Job Title" field. The model generally exhibits poor performance when extracting these types of attributes, with the notable exception of the "Police Number". The model shows exceptional accuracy in extracting the "Police Number", which can be attributed to the fact that this attribute is typically represented as a distinct 6-digit numerical sequence. Such a pattern is relatively straightforward for the model to identify and recognize due to its unique numerical structure. Other fields that share similar characteristics, such as "ID Number", "Phone Number", and "Report Time", also demonstrate higher extraction performance.

Table 3: Statistics of the number of each field in the training set

Fields	Counts	Fields	Counts
Police Number	243	Registered Address	646
Job Title	445	Current Address	742
Name	2386	Marital Status	221
ID Number	1718	Phone Number	1533
Native Place	111	Dispute Party	1485
Ethnicity	432	Report Time	811

However, for the “Dispute Party” field, despite having a relatively ample amount of data in the training set, the model’s performance in extracting this information remains relatively poor. During this verification process, we discovered that extracting information for the “Dispute Party” field is not a straightforward task; rather, it involves a degree of reasoning and inference. For example, in the sentence “Alice insists on selling the house, but Bob refuses to leave,” the fact that Alice and Bob are parties to a dispute is not explicitly stated. However, it can be logically inferred that they are in conflict over the decision to sell the house. This requirement for reasoning applies similarly to the “Marital Status” field. For a detailed analysis of this issue, please refer to the case study in [Appendix B](#). In essence, these fields require the model to go beyond simple NER tasks and exhibit more advanced reasoning capabilities, which is why their performance is noticeably lower compared to fields involving more straightforward information extraction.

4.4 Ablation Study

We performed fine-tuning on the Qwen2-7B-Instruct model using both the training sets derived from the Generated Dataset and Verified Dataset as mentioned in [Section 4.2](#), employing the same parameter settings for consistency. The performance of these fine-tuned models was evaluated using the verified test set, with the results presented in [Table 4](#).

Table 4: F1 scores comparison of Qwen2 fine-tuned on two datasets. Bold “Overall” value (84.75) reflects the optimal F1 achieved using the verified training set, showing a 9.39-point improvement over the generated training set (75.36)

Fields	Generated train set	Verified train set	Improvement
Police Number	89.32	94.71	5.39
Job Title	41.10	42.91	1.80
Name	82.91	91.00	8.09
ID Number	82.19	91.24	9.04
Native Place	35.44	49.23	13.79
Ethnicity	66.88	73.39	6.51
Registered Address	64.60	77.11	12.51
Current Address	59.63	72.57	12.94
Marital Status	28.39	36.41	8.01
Phone Number	76.90	87.91	11.01
Dispute Party	56.96	66.00	9.04
Report Time	92.29	97.38	5.09
Overall	75.36	84.75	9.39

The results clearly indicate that the Qwen2 model fine-tuned on the manually verified dataset exhibits a significant improvement in performance, with an overall increase of approximately 10% in evaluation metrics. This finding underscores the fact that our task presents certain complexities that even powerful LLMs may struggle to fully address. Moreover, it highlights that datasets generated by these models are not infallible and can contain errors or inconsistencies. Therefore, the manual verification and correction of training data are critical steps in enhancing the accuracy and reliability of information extraction models, particularly when these models are applied in practical, real-world scenarios.

5 Conclusion

This paper proposes the TIPS, which uses a template-based synthetic dataset to perform LoRA fine-tuning on large language models and conducts joint extraction tasks based on the carefully designed extraction scheme. It enables complex multi-task information extraction in real-world scenarios and offers a viable solution for efficient information extraction in low-resource environments. Experiments show that, after fine-tuning with limited public security data, the model's information extraction capability significantly improved. Moreover, manual validation showed a 9.39% increase in overall F1 score, with GLM-4 achieving the highest performance at 87.14%. For basic fields like police ID numbers, the F1 score reached 94.71%, while low-resource fields such as job titles remained lower at 42.91%, underscoring the need for further optimization in these fields.

The experimental results highlight the strong potential of LLMs in public security information extraction. Fine-tuning with domain-specific data significantly improves accuracy, adaptability, and robustness, providing crucial technical support for intelligent public security systems. The proposed method has been successfully deployed in a city's public security big data system, enabling entity, attribute, and relationship extraction from multi-source data such as alarm records and public opinion information. These results have been applied to public security incident analysis and knowledge graph construction. Future research will focus on enhancing data diversity (e.g., low-resource fields like Job Title with only 445 instances), extending TIPS to dynamic event reasoning (e.g., predicting escalation risks from dispute patterns), and refining prompt templates to improve implicit relationship reasoning.

Acknowledgement: The authors would like to express their gratitude to the editors and reviewers for their detailed review and insightful advice.

Funding Statement: This work is supported by the National Key Research and Development Program of China (2021YFC3300500) and the Ministry of Public Security Technical Research Program (2023JSZ01).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yue Liu; data collection: Yue Liu, Qinglang Guo; experiments: Yue Liu, Qinglang Guo, Chunyao Yang; analysis and interpretation of results: Yue Liu; draft manuscript preparation: Yue Liu, Qinglang Guo, Chunyao Yang; supervision: Yong Liao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Nomenclature

NLP	Natural Language Processing
IE	Information Extraction
LLM	Large Language Model
LoRA	Low-Rank Adaptation
NER	Named Entity Recognition

Appendix A. Dataset Sample Display

Fig. A1 presents two examples from the dataset used in this study, including the original Chinese version and the translated English version. Each example consists of a sample and its corresponding labels. The model generates the labels based on the samples, following instructions and an extraction schema, which are then manually verified for accuracy.

Chinese	English
<p>2022年8月5日17时30分，报案人陈华（男，身份证号510112197610121234，住址：成都市青羊区人民南路333号，联系电话13898761234）称他在一家购物网站上购买了一台电视，花费了4500元人民币，但商品送到后发现存在质量问题。他要求退货或维修，但卖家拒绝了他的请求，卖家表示问题是由物流公司造成的。双方在产品质量和责任问题上产生了争议，希望警方协助解决此购物纠纷。</p> <p>报警人： - 姓名：陈华 - 身份证号：510112197610121234 - 户籍地址：成都市青羊区人民南路333号 - 手机号码：13898761234 - 纠纷对象：卖家</p> <p>报警时间：2022-08-05T17:30:00</p> <p>纠纷当事人1： - 姓名：陈华 - 身份证号：510112197610121234 - 户籍地址：成都市青羊区人民南路333号 - 手机号码：13898761234 - 纠纷对象：卖家</p> <p>纠纷当事人2： - 姓名：卖家 - 纠纷对象：陈华</p>	<p>On August 5, 2022, at 5:30 PM, the reporter Chen Hua (male, ID number: 510112197610121234, address: 333 Renmin South Road, Qingyang District, Chengdu City, contact number: 13898761234) reported that he purchased a television from an online shopping website for 4500 RMB. Upon delivery, he discovered quality issues with the product. He requested a refund or repair, but the seller refused, claiming that the issue was caused by the logistics company. Both parties have disputes over the product quality and responsibility, and the complainant seeks police assistance to resolve this shopping dispute.</p> <p>Reporter: - Name: Chen Hua - ID Number: 510112197610121234 - Registered Address: 333 Renmin South Road, Qingyang District, Chengdu City - Phone Number: 13898761234 - Dispute Party: Seller</p> <p>Report Time: 2022-08-05T17:30:00</p> <p>Dispute Party 1: - Name: Chen Hua - ID Number: 510112197610121234 - Registered Address: 333 Renmin South Road, Qingyang District, Chengdu City - Phone Number: 13898761234 - Dispute Party: Seller</p> <p>Dispute Party 2: - Name: Seller - Dispute Party: Chen Hua</p>
Chinese	English
<p>出警民警：江永昶，警号：006604。带领2名辅警（姓名：刘军益、金鑫）。经了解，报警人章国强（342529199011082218）称其母亲叶爱五（342529196511212222）在小区内被一个大约十岁的小孩（身份信息不详）骑车时不小心撞到了，现其已带其母亲到医院检查，其爱人（身份信息未提供，18326069285）已到小区查找监控。求助民警帮助联系对方。民警现已帮助其查找监控处理。</p> <p>报警人： - 姓名：章国强 - 身份证号：342529199011082218</p> <p>接警人1： - 姓名：江永昶 - 警号：006604 - 职务：民警</p> <p>接警人2： - 姓名：刘军益 - 职务：辅警</p> <p>接警人3： - 姓名：金鑫 - 职务：辅警</p> <p>纠纷当事人1： - 姓名：叶爱五 - 身份证号：342529196511212222 - 纠纷对象：一个大约十岁的小孩</p> <p>纠纷当事人2： - 姓名：一个大约十岁的小孩 - 纠纷对象：叶爱五</p> <p>其他相关人： - 姓名：章国强的爱人 - 手机号码：18326069285</p>	<p>Responding police officer: Jiang Yongchang, Badge Number: 006604. Accompanied by two auxiliary police officers (names: Liu Junyi, Jin Xin). According to the reporter Zhang Guoqiang (ID: 342529199011082218), his mother Ye Aiwu (ID: 342529196511212222) was accidentally hit by a child about ten years old (identity unknown) riding a bicycle in the community. Zhang has taken his mother to the hospital for examination, and his spouse (identity not provided, contact number: 18326069285) has gone to the community to check surveillance footage. The police have helped retrieve the surveillance footage and handle the matter.</p> <p>Reporter: - Name: Zhang Guoqiang - ID Number: 342529199011082218</p> <p>Responding Officer 1: - Name: Jiang Yongchang - Badge Number: 006604 - Role: Police Officer</p> <p>Responding Officer 2: - Name: Liu Junyi - Role: Auxiliary Police</p> <p>Responding Officer 3: - Name: Jin Xin - Role: Auxiliary Police</p> <p>Dispute Party 1: - Name: Ye Aiwu - ID Number: 342529196511212222 - Dispute Object: A child about ten years old</p> <p>Dispute Party 2: - Name: A child about ten years old - Dispute Object: Ye Aiwu</p> <p>Other Related Persons: - Name: Zhang Guoqiang's spouse - Phone Number: 18326069285</p>

Figure A1: Dataset samples and corresponding labels

Appendix B. Case Study

The evaluation results presented in this appendix were obtained using our fine-tuned ChatGLM4-9B model. Here is the original police report text, label, and prediction results.

At 12:10 PM on January 8, 2023, Li Lei (male, ID number: 610202199006188888, address: No. 888, Zhujiang New Town, Tianhe District, Guangzhou City, contact number: 13888889999) reported that he had purchased a mobile phone and discovered that the WeChat Pay on the device had been maliciously tampered with, leading to the leakage of his personal information. He provided the download link for the app: <http://www.wechat.com>. Zhao Li (female, ID number: 330203199504087890, address: No. 36, Renmin Avenue, Yuexiu District, Guangzhou City, contact number: 13987654321) is one of the developers of the involved app and has been listed as a suspect. This is a case of an information leakage dispute related to an app.

```
{
  "Reporter": [
    {
      "Name": "Li Lei",
      "IDNumber": "610202199006188888",
      "NativePlace": null,
      "Ethnicity": null,
      "RegisteredAddress": null,
      "CurrentAddress": "No. 888, Zhujiang New Town, Tianhe District, Guangzhou",
      "MaritalStatus": null,
      "PhoneNumber": "13888889999",
      "DisputeTargets": "Zhao Li"
    }
  ],
  "DisputeParties": [
    {
      "Name": "Li Lei",
      "IDNumber": "610202199006188888",
      "NativePlace": null,
      "Ethnicity": null,
      "RegisteredAddress": null,
      "CurrentAddress": "No. 888, Zhujiang New Town, Tianhe District, Guangzhou",
      "MaritalStatus": null,
      "PhoneNumber": "13888889999",
      "DisputeTargets": "Zhao Li"
    },
    {
      "Name": "Zhao Li",
      "IDNumber": "330203199504087890",
      "NativePlace": null,
      "Ethnicity": null,
      "RegisteredAddress": null,
      "CurrentAddress": "No. 36, Renmin Avenue, Yuexiu District, Guangzhou",
      "MaritalStatus": null,
      "PhoneNumber": "13987654321",
      "DisputeTargets": "Li Lei"
    }
  ],
  "RelatedPersons": [],
  "PoliceOfficer": [],
  "ReportTime": "2023-01-08T12:10:00"
}
```

Listing A1: Label


```

{
  "Reporter": [
    {
      "Name": "Li Lei",
      "IDNumber": "610202199006188888",
      "NativePlace": null,
      "Ethnicity": null,
      "RegisteredAddress": null,
      "CurrentAddress": "No. 888, Zhujiang New Town, Tianhe District, Guangzhou",
      "MaritalStatus": null,
      "PhoneNumber": "13888889999",
      "DisputeTargets": null
    }
  ],
  "DisputeParties": [],
  "RelatedPersons": [
    {
      "Name": "Zhao Li",
      "IDNumber": "330203199504087890",
      "NativePlace": null,
      "Ethnicity": null,
      "RegisteredAddress": null,
      "CurrentAddress": "No. 36, Renmin Avenue, Yuexiu District, Guangzhou",
      "MaritalStatus": null,
      "PhoneNumber": "13987654321",
      "DisputeTargets": null
    }
  ],
  "PoliceOfficer": [],
  "ReportTime": "2023-01-08T12:10:00"
}

```

Listing A2: Prediction Results of ChatGLM-4-9B model

The evaluation results demonstrate that our fine-tuned ChatGLM4-9B model is highly effective at extracting explicit personal information from the police report. Key details such as names, ID numbers, addresses, and phone numbers were accurately extracted, highlighting the model's proficiency in entity extraction. This level of performance aligns with traditional information extraction tasks.

However, our task goes beyond traditional information extraction tasks, including both entity extraction and the more complex relationship and event extraction. While traditional systems focus on identifying entities and their relationships, our task requires a deeper level of reasoning. Specifically, it involves understanding and classifying the roles of the entities (e.g., identifying who the reporter is, who the dispute parties are), inferring their relationships, and interpreting how these relationships evolve in the context of the situation. This represents a significant shift from traditional methods, which typically operate on surface-level entity relationships and event extraction without necessarily interpreting the broader context.

For example, in the provided police report, Zhao Li is misclassified as a "Related Person" rather than a "Dispute Party." This misclassification arises from the model's inability to infer the deeper relationships between the entities. While traditional information extraction systems can identify that Zhao Li is mentioned in the report, they might not capture the implicit relationship that Zhao Li is a key figure in the dispute. This failure to properly classify the dispute parties leads to errors in identifying the correct entities involved in the event, which in turn distorts the overall understanding of the situation.

Thus, the challenge we face is not merely one of extracting relationships between entities but of reasoning about these relationships in the context of a broader narrative. Our task involves the complex interplay of entity classification, relationship inference, and contextual understanding. While traditional

information extraction may rely on pattern matching and rule-based classification, our task requires deeper comprehension and reasoning, as it involves understanding not only the entities but also the underlying relationships that shape their roles within the context of the event.

References

1. Cowie J, Lehnert W. Information extraction. *Commun ACM*. 1996 Jan;39(1):80–91. doi:10.1145/234173.234209.
2. Adnan K, Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. *J Big Data*. 2019 Oct;6(1):1. doi:10.1186/s40537-019-0254-8.
3. Benkassioui B, Kharmoum N, Hadi MY, Ezziyyani M. NLP methods' information extraction for textual data: an analytical study. In: Kacprzyk J, Ezziyyani M, Balas VE, editors. *International Conference on Advanced Intelligent Systems for Sustainable Development*; 2023; Cham: Springer Nature Switzerland. p. 515–27.
4. Yang Y, Wu Z, Yang Y, Lian S, Guo F, Wang Z. A survey of information extraction based on deep learning. *Appl Sci*. 2022;12(19):9691. doi:10.3390/app12199691.
5. Li Q, Li J, Sheng J, Cui S, Wu J, Hei Y, et al. A survey on deep learning event extraction: approaches and applications. *IEEE Trans Neural Netw Learn Syst*. 2024;35(5):6301–21. doi:10.1109/TNNLS.2022.3213168.
6. Deng S, Ma Y, Zhang N, Cao Y, Hooi B. Information extraction in low-resource scenarios: survey and perspective. *arXiv:2202.08063*. 2023.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, et al., editors. *Advances in neural information processing systems*. Curran Associates, Inc.; 2017. Vol. 30. [cited 2024 Dec 30] Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
8. Hedderich MA, Lange L, Adel H, Strötgen J, Klakow D. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv:2010.12309*. 2020.
9. Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. *Computer*. 2004;37(4):50–6. doi:10.1109/MC.2004.1297301.
10. Yang Z, Liu Y, Ouyang C. Causal intervention-based few-shot named entity recognition. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics: EMNLP 2023*; 2023; Singapore: Association for Computational Linguistics. p. 15635–46.
11. Amalvy A, Labatut V, Dufour R. Learning to rank context for named entity recognition using a synthetic dataset. In: Bouamor H, Pino J, Bali K, editors. *Proceedings of the 2023 Conference on Hods In Natural Language Processing*; 2023; Singapore: Association for Computational Linguistics. p. 10372–82.
12. Xu B, Wang Q, Lyu Y, Dai D, Zhang Y, Mao S. S2ynRE: two-stage self-training with synthetic data for low-resource relation extraction. In: Rogers A, Boyd-Graber J, Okazaki N, editors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 8186–207.
13. Zhou W, Zhang S, Naumann T, Chen M, Poon H. Continual contrastive finetuning improves low-resource relation extraction. In: Rogers A, Boyd-Graber J, Okazaki N, editors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 13249–63.
14. Gao J, Yu C, Wang W, Zhao H, Xu R. Mask-then-fill: a flexible and effective data augmentation framework for event extraction. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Findings of the Association for Computational Linguistics: EMNLP 2022*; 2022; Abu Dhabi, Anited Arab Emirates: Association for Computational Linguistics. p. 4537–44.
15. Lu D, Ran S, Tetreault J, Jaimes A. Event extraction as question generation and answering. In: Rogers A, Boyd-Graber J, Okazaki N, editors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 1666–88.

16. Ye D, Lin Y, Li P, Sun M. Packed levitated marker for entity and relation extraction. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022; Dublin, Ireland: Association for Computational Linguistics. p. 4904–17.
17. Zheng Y, Hao A, Luu AT. Jointprop: joint semi-supervised learning for entity and relation extraction with heterogeneous graph-based propagation. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023; Toronto, ON, Canada: Association for Computational Linguistics. p. 14541–55.
18. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021.
19. Li Y, Gai K, Qiu L, Qiu M, Zhao H. Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Inf Sci.* 2017;387(5):103–15. doi:10.1016/j.ins.2016.09.005.
20. Tian Z, Li X. Research on Chinese event detection method based on BERT-CRF model. *Comput Eng Appl.* 2021;57(11):135–9.
21. Ku CH, Iriberri A, Leroy G. Natural language processing and e-government: crime information extraction from heterogeneous data sources. In: Proceedings of the 2008 International Conference on Digital Government Research; 2008; California: Digital Government Society of North America. p. 162–70.
22. Chau M, Xu JJ, Chen H. Extracting meaningful entities from police narrative reports. In: Proceedings of the 2002 Annual National Conference on Digital Government Research; 2002; California: Digital Government Society of North America. p. 1–5.
23. Ku CH, Iriberri A, Leroy G. Crime information extraction from police and witness narrative reports. In: 2008 IEEE Conference on Technologies for Homeland Security; 2008; New York City, NY, USA: IEEE. p. 193–8.
24. Deng Q, Xie S, Zeng D, Zheng F, Cheng C, Peng L. An event extraction method for public security. *J Chin Inf Process.* 2022;36(9):93–101.
25. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018.
26. Dey R, Salem FM. Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS); 2017; New York City, NY, USA: IEEE. p. 1597–600.
27. Bale TL, Vale WW. CRF and CRF receptors: role in stress responsivity and other behaviors. *Annu Rev Pharmacol Toxicol.* 2004;44(1):525–57. doi:10.1146/annurev.pharmtox.44.101802.121410.
28. Zhang X, Yang W, Hu Y, Liu J. DMCNN: dual-domain multi-scale convolutional neural network for compression artifacts removal. In: 2018 25th IEEE International Conference on Image Processing (ICIP); 2018; New York City, NY, USA: IEEE. p. 390–4.
29. Cui Y, Wang J, Yan S, Tao Z. Automatic key information extraction of police records based on deep learning. *Big Data Res.* 2022;8(6):127–42.
30. Zhao C, Xie S, Zeng D, Zheng F, Cheng C, Peng L. Combination of pre-trained language model and label dependency for relation extraction. *J Chin Inf Process.* 2022;36(1):75–82.
31. Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, et al. Harnessing the power of llms in practice: a survey on chatgpt and beyond. *ACM Trans Knowl Discov Data.* 2024;18(6):1–32. doi:10.1145/3649506.
32. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv:2303.18223. 2023.
33. Xiao C, Xu SX, Zhang K, Wang Y, Xia L. Evaluating reading comprehension exercises generated by LLMs: a showcase of ChatGPT in education applications. In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023); 2023; Stroudsburg, PA, USA: Association for Computational Linguistics. p. 610–25.
34. Kolasani S. Optimizing natural language processing, large language models (LLMs) for efficient customer service, and hyper-personalization to enable sustainable growth and revenue. *Trans Latest Trends Artif Intell.* 2023;4(4).
35. Zhu Y, Yuan H, Wang S, Liu J, Liu W, Deng C, et al. Large language models for information retrieval: a survey. arXiv:2308.07107. 2023.
36. Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. arXiv:2303.15647. 2023.

37. Zaken EB, Ravfogel S, Goldberg Y. Bitfit: simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv:2106.10199. 2021.
38. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691. 2021.
39. Su Y, Wang X, Qin Y, Chan CM, Lin Y, Wang H, et al. On transferability of prompt tuning for natural language processing. arXiv:2111.06719. 2021.
40. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al. Qwen2 technical report. arXiv:2407.10671. 2024.
41. T. GLM, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. arXiv:2406.12793. 2024.
42. AI@Meta. Llama 3 model card. 2024 [cited 2024 Jul 11]. Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
43. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas DDL, et al. Mistral 7B. arXiv:2310.06825. 2023.
44. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002; Stroudsburg, PA, USA: Association for Computational Linguistics. p. 311–8.
45. Lin CY. Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. p. 74–81.
46. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv:2010.16061. 2020.
47. Zheng Y, Zhang R, Zhang J, Ye Y, Luo Z, Feng Z, et al. LlamaFactory: unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations); 2024; Bangkok, Thailand: Association for Computational Linguistics.
48. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101. 2017.