Improving unified information extraction in Chinese mental health domain with instruction-tuned LLMs and type-verification component

Improving unified information extraction in Chinese mental health domain with instruction-tuned LLMs and type-verification component

Zijie Cai[a,b], Hui Fang[c,*], Jianhua Liu[a,b], Ge Xu[c], Yunfei Long[d], Yin Guan[c], Tianci Ke[a,b]

[a]*School of Computer Science and Mathematics, Fujian University of Technology, 350118, Fuzhou, China*
[b]*Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, 350118, China*
[c]*School of Computer and Big Data, Minjiang University, 350108, Fuzhou, China*
[d]*School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, Colchester, UK*

Abstract

Background: Extracting psychological counseling help-seeker information from unstructured text is crucial for providing effective mental health support. This task involves identifying personal emotions, psychological states, and underlying psychological issues but faces significant challenges. These challenges include the sensitivity of mental health data, the lack of Chinese instruction datasets, and the difficulties large language models (LLMs) encounter with complex natural language understanding tasks..
Objective: This study aims to address these challenges by developing a unified information extraction framework for Chinese mental health texts. Specifically, it leverages instruction-tuned LLMs and incorporates a novel type-verification (TV) component to improve performance while minimizing computational demands.
Methods: We first constructed a Chinese mental health domain instruction dataset for mental health information extraction using synthetic data generated by ChatGPT, guided by psychology experts. This dataset includes self-reported statements from psychological counseling help-seekers, capturing their personal situations, emotions, thoughts, and experiences. Subsequently, we fine-tuned open-source LLMs on this dataset to perform named entity recognition, relation extraction, and event extraction. To address errors and omissions in the extracted information, we introduced a type-verification component. This component employs a lightweight model with significantly fewer parameters to verify the extracted types. The verification results were then fed back into LLMs for further refinement.
Results: Experimental results demonstrate that our framework achieves outstanding performance in mental health information extraction. The type-verification component significantly enhances extraction accuracy while reducing computational resource requirements through the use of a lightweight model. By combining robust instruction-tuned LLMs with an efficient type-verification component, our approach delivers exceptional results.
Conclusion: This study presents a novel and efficient framework for tackling the challenges of mental health information extraction in Chinese texts. By integrating instruction-tuned LLMs with a lightweight type-verification component, our approach significantly improves extraction accuracy and computational efficiency. This framework holds promise for supporting scalable, automated mental health support systems, advancing both research and practical applications in the mental health domain.

Keywords: Information extraction, Large language models, Mental health, Type-verification

1. Introduction

Recently, mental health issues have received increasing attention[1], prompting numerous studies that integrate Large language models(LLMs) into the mental health domain. Xu et al. [2] present a comprehensive evaluation of multiple LLMs on various mental health prediction tasks using online data and summarize their findings into actionable guidelines to improve the ability of LLMs to predict mental health through online data. Cabrera et al. [3] analyzed the bioethical dilemmas associated with the use of chatbots in mental health and suggested that chatbots should complement, rather than replace, the treatment programs provided by human professionals. PsyQA [4] is a Chinese psychological health support dataset comprising question-and-answer pairs. Building on PsyQA, Qiu et al. [5] constructed a large-scale multi-round dialogue corpus that closely resembles real-life scenarios using the proposed SMILE method. They subsequently developed MeChat, a general-purpose model that supports Chinese mental health applications, by adjusting the ChatGLM-6B foundation model.

Mental health research has traditionally focused on natural language generation (NLG) tasks, while natural language understanding (NLU) tasks are gaining less attention. In psychological counseling, extracting meaningful information from unstructured text is essential for effective mental health support. This process involves identifying individuals' emotions, psychological states, and underlying psychological issues, which

*Corresponding author
Email addresses: zhijie2961966@qq.com (Zijie Cai), fh_mju@126.com (Hui Fang), jhliu@fjnu.edu.cn (Jianhua Liu), xuge@mju.edu.cn (Ge Xu), yl20051@essex.ac.uk (Yunfei Long), niynaug@foxmail.com (Yin Guan), tiancike@foxmail.com (Tianci Ke)

requires the expertise of trained professionals to ensure accurate interpretation. As a result, NLU tasks in this domain impose higher demands on both data construction and model development to achieve reliable and meaningful outcomes.

However, the field faces several challenges. First, research on text-based mental health counseling is significantly constrained by the scarcity of relevant corpora, particularly Chinese instructional datasets. Second, while large models perform adequately on simpler natural language understanding (NLU) tasks, they struggle with more complex ones, such as challenging sequence labeling tasks like named entity recognition (NER)[6]. These limitations hinder the development of robust AI-driven mental health support systems, emphasizing the need for more comprehensive datasets and advanced model architectures.[7].

To address these challenges, we constructed a Chinese instruction dataset designed to unify information extraction in the mental health domain. The dataset was generated by prompting ChatGPT to generate synthetic data under the guidance of a psychology experts. It includes self-statements of help seekers in psychological counseling, capturing their situations, emotions, thoughts, experiences, and other critical information for extraction. With this dataset, we instruction-tuned open-source LLMs to perform information extraction tasks in the mental health domain. However, we observed that LLMs often make errors and omissions when identifying the correct types of information, leading to inaccuracies in the extracted data and ultimately affecting the overall performance of information extraction. To enhance the performance of LLMs on this task, we propose a type-verification (TV) method tailored to the mental health domain. This method introduces a lightweight verification model with significantly fewer parameters than LLMs for training. The verification model describes the process as a classification problem and the instructions as input and output of the classification results. These verification results are then sent back to LLMs for further processing to improve the accuracy and reliability of the information extraction.

Our contributions can be summarized as follows:

- We created unified information extraction instruction datasets specifically tailored to the mental health domain. These datasets encompass tasks such as named entity recognition, relation extraction, and event extraction.
- We utilized various LLMs to develop a unified information extraction model for mental health, improving robustness and performance through multi-Instruction-tuning and data augmentation.
- We introduced a novel approach leveraging small-parameter models for type-verification in collaboration with LLMs. Experimental results demonstrate that this method significantly enhances the performance of LLMs in the extraction of mental health information, validating the efficacy of small-parameter models.
- We constructed a type verification training dataset from the original instruction datasets and designed a straightforward method to generate positive and negative samples.

2. Related work

The related works relevant to our paper focus on three key aspects: unifying information extraction, instruction-tuning, and improving the information extraction capabilities of LLMs.

2.1. Unify information extraction

Traditional information extraction methods [8, 9, 10] typically rely on designing specific architectures for different information extraction tasks [11]. This approach makes it challenging to extend these systems to new tasks, significantly limiting their versatility and practical applications. To address these issues, Lu et al. [12] proposed Universal Information Extraction, which aims to handle multiple information extraction tasks using a common model. It encodes various extraction structures in a unified manner through a structured extraction language and captures universal extraction capabilities with a large-scale pre-trained text-to-structure model. However, it requires fine-tuning for different downstream tasks, leading to suboptimal performance in low-resource scenarios.

To overcome this limitation, Lou et al. [13] introduced Unified Semantic Matching (USM), which decouples information extraction into two fundamental tasks by uniformly encoding tag patterns and text within a shared semantic representation. Despite its innovation, USM's reliance on word-level semantic matching results in a significant increase in both training and inference times. Wang et al. [14] proposed InstructUIE, which leverages instructive guidance to steer pre-trained large models toward specific tasks. This approach facilitates the efficient and adaptive generation of target structures. However, InstructUIE primarily focuses on the general domain and requires substantial computational resources to fine-tune the model with full parameters.

2.2. Instruction-tuning

Instruction-tuning is the process of further training large language models on instruction datasets composed of instruction-output pairs [15]. Its uniqueness lies in the structure of these datasets, which consist of human instructions and their corresponding expected outputs. This structure enhances the model's ability to understand and follow instructions. Consequently, research on instruction-tuning primarily focuses on the construction of instruction datasets. Wang et al. [16] proposed SELF-INSTRUCT, a method for generating the instruction datasets needed for instruction-tuning by guiding pre-trained language models. However, recent research on instruction-tuning predominantly focuses on non-information extraction tasks [17], such as question answering and text classification. Although some studies [18, 19, 14] address most information extraction tasks, they are oriented toward the needs of the general domain.

2.3. Improving information extraction of LLMs

Recently, large language models (LLMs) have demonstrated tremendous capabilities in solving a variety of natural language

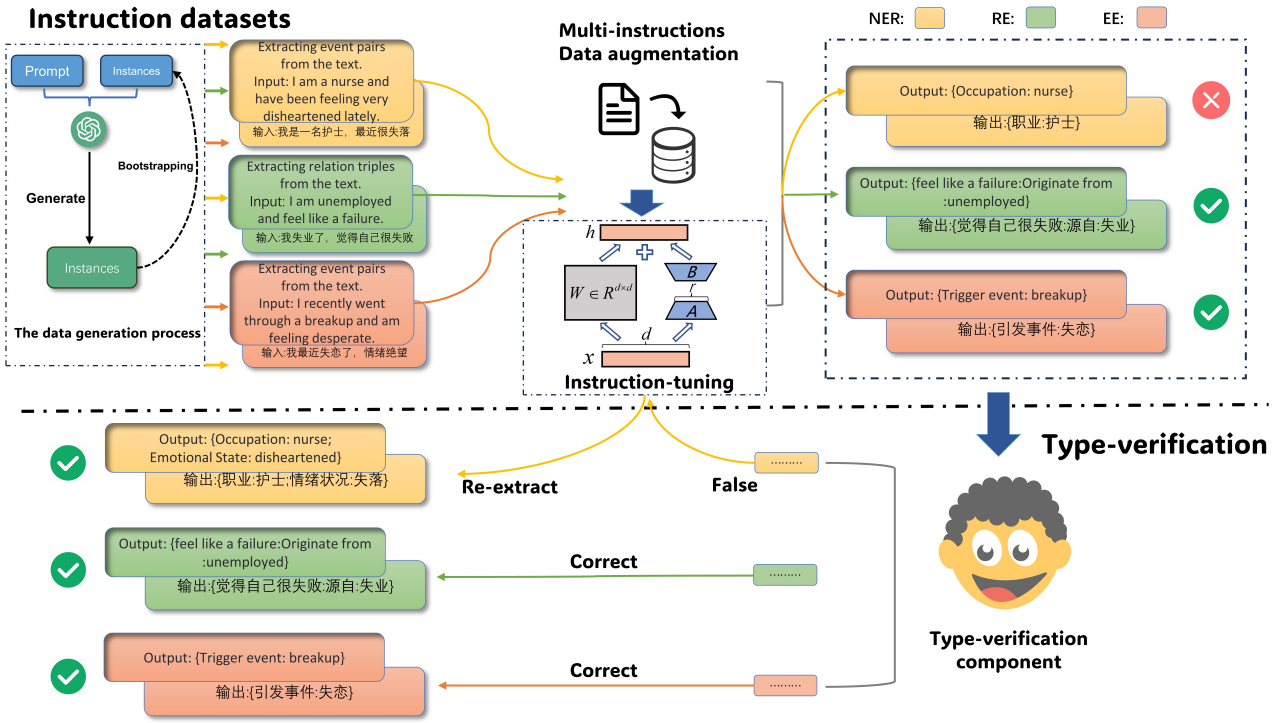

Fig. 1: The overall framework. The upper part illustrates the construction of the mental health instruction dataset and the instruction-tuning techniques introduced in Section 3.1, and the lower part details the type-verification component and implementation process described in Section 3.2.

tasks, showcasing strong generalization abilities [20]. However, researchers have identified several limitations and drawbacks when applying LLMs to information extraction tasks, including hallucination, performance degradation in specific domains, and overprediction of NULL cases, among others. To address these issues and enhance information extraction performance, Wan et al. [21] conducted relation extraction on LLMs via in-context learning (ICL) and proposed GPT-RE. This method employs task-aware demonstration retrieval to mitigate the strong tendency to misclassify NULL examples as other predefined labels. However, it is constrained by the access restrictions of GPT-3, and the recall for NULL cases remains inferior to fully-supervised baselines. Wang et al. [22] introduced GPT-NER to bridge the gap between named entity recognition and LLMs, incorporating a self-verification strategy to further improve performance. Nevertheless, all steps in this method must be executed separately for each entity type and sentence, significantly increasing the usage of LLMs' API.

In the medical domain, Hu et al. [23] proposed zero-shot information extraction from radiological reports using ChatGPT, incorporating prior medical knowledge into the prompt template to reduce extraction errors. However, its performance is subpar in complex tasks and cannot be optimized for specific tasks through fine-tuning, which limits its effectiveness in certain specialized domains or tasks. Rohanian et al. [24] explored the potential of instruction-tuning for biomedical language processing, applying this technique to two general LLMs of substantial scale. This approach improved the performance of biomedical and clinical Named Entity Recognition (NER), Relation Extraction (RE), and Medical Natural Language Inference (NLI). Notably, however, these models did not outperform specialized models and failed to further enhance information extraction performance. Datta et al. [25] leveraged large language models to extract granular eligibility criteria for diverse diseases from free-text clinical trial protocol documents, building a generalizable information extraction system. However, the system struggles to handle certain sophisticated criteria conditions.

2.4. Data augmentation

Data augmentation aims to generate intelligible and varied additional instances while preserving semantic congruity. In certain scenarios, it is particularly important due to the limited richness of available training data. Zhang et al. [26] introduced three simple yet effective strategies using LLMs to augment original training instances, addressing resource shortages in the Chinese dialogue-level dependency parsing task. Yuan et al. [27] proposed an innovative privacy-aware data augmentation approach that leverages the advantages of LLMs while ensuring the security and confidentiality of sensitive patient data. Strelcenia et al. [28] investigated various data augmentation techniques to address the problem of imbalanced data, improving classification performance in credit card fraud detection by introducing a novel data augmentation model. However, none of the aforementioned methods leverage data augmentation to enhance the robustness of LLMs.

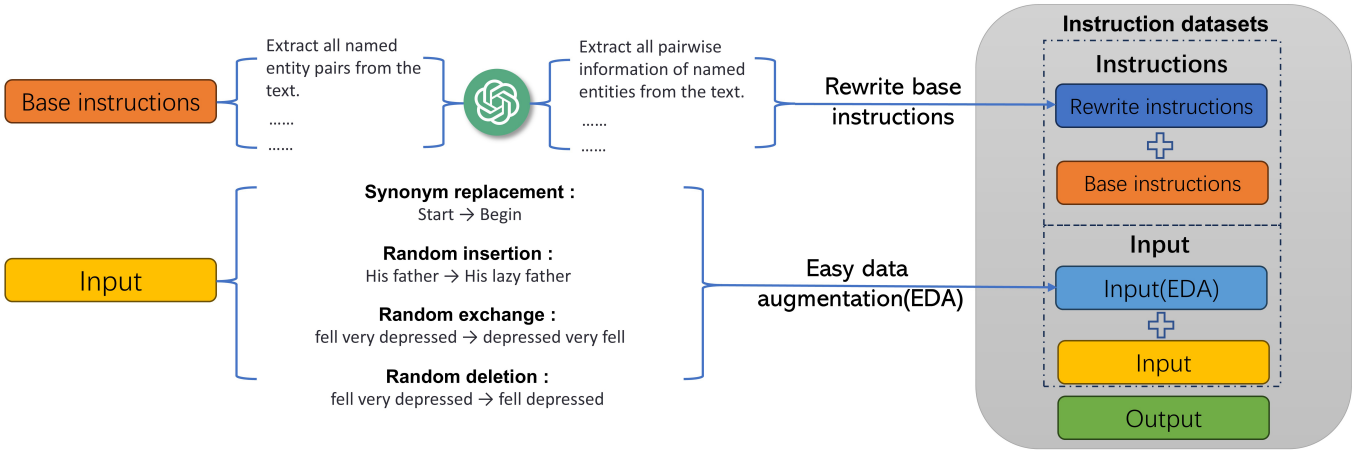

Fig. 2: Multi-instructions and data augmentation for instruction datasets.

3. Methodology

In this section, we first review the construction of mental health instruction datasets and the techniques used in instruction-tuning in Section 3.1. Next, we provide a detailed introduction to the implementation process and application of type-verification in Section 3.2. The overall framework is shown in Fig. 1 and the abbreviations we used are shown in Table 1.

Table 1: Abbreviations.

LLMs	Large language models
IE/ie	Information extraction
UIE	Unify information extraction
TV/tv	Type-verification
DA	Data augmentation
SFT	Supervised Fine-tuning
FT	Fine-tuning

3.1. Construction of instruction datasets

3.1.1. Definition and Construction

For a given set of task instructions $\{I_t\}$, each instruction defines its related tasks. Each task contains one or more input and output instances (x, y). Given a task instruction I_t and an input instance x, a model M is expected to produce an output instance y: $M(I_t, x) = y$. A complete instruction datasets I consists of task instructions $\{I_t\}$, input instances x and output instances y: $I = \{I_t, x, y\}$.

To address the poor performance of ChatGPT in information extraction tasks within specific domains, as well as the associated issues of privacy leakage[29]. In this work, we utilize ChatGPT to generate instances of each task. This method can generate entire training datasets with only a few samples, without the need for manual labeling tools or other natural language processing tools, thus solving the common problems of difficult data acquisition and high data labeling costs in the mental health domain. The data generation process is shown in the upper part of Fig. 1, it requires a few initial instances with x and y to generate a set of new instances. The instances generation process is then iteratively repeated with these new instances to construct the complete dataset with instances x and y. We present all types and quantity statistics of three information extraction tasks in Appendix A.

3.1.2. Multi-instructions and data augmentation

After obtaining the input and output examples (x, y) of the instruction datasets as described in Section 3.1.1, the corresponding task instruction I_t is required. At this stage, we proposed a multi-instruction learning method to enhance the reliability and generalization of the output results of LLMs. Specifically, we manually designed eight different instructions for three different tasks and utilized ChatGPT to rewrite these base instructions. Therefore, each task includes 16 different instructions, which describe the corresponding tasks but differ in expression and complexity.

Spelling mistakes are common in real-world texts and can usually be corrected automatically using context information. However, even ChatGPT exhibits limitations compared to humans in the task of detecting grammar, syntax, and spelling mistakes in sentences[30]. Moreover, due to Chinese character-level constraints, the performance of LLMs in detecting and correcting spelling mistakes in Chinese sentences is often unsatisfactory[31]. To this end, we utilize easy data augmentation[32] to introduce noise, thereby promoting the model's ability to handle imperfect or inaccurate input and enhancing the robustness of LLMs. To this end, we conduct the data augmentation at the character level and word level based on the original training set. Specifically, we perform disturbing operations on the input part of each input instance after removing stop words, such as synonym replacement, random insertion, random exchange, and random deletion. The proportion of changed words in each input instance is 10%. Then, the repeated samples are removed, and the original data and the augmented data are combined to be used as the data of the training model. We show the process described above in Fig. 2.

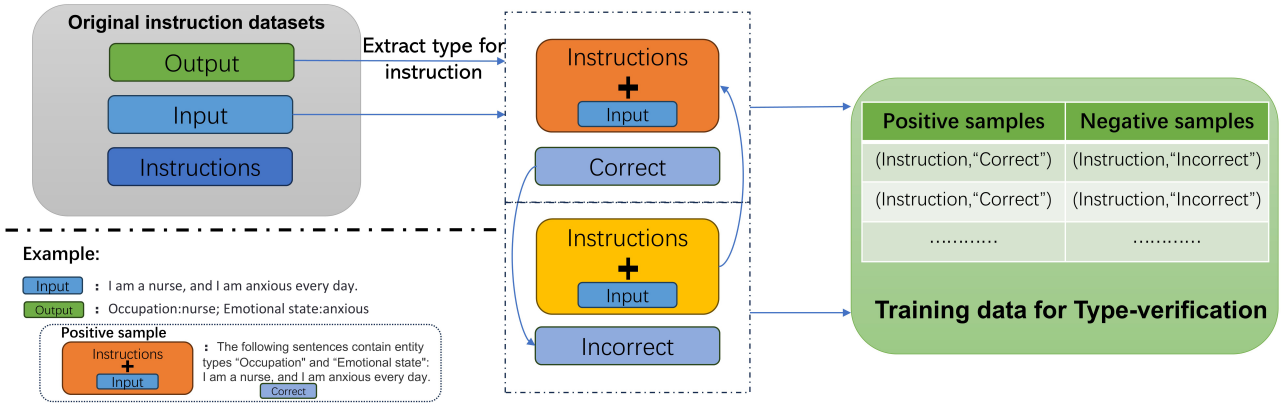

Fig. 3: Data acquisition process of type-verification component.

3.1.3. Instruction-tuning

Traditional fine-tuning requires training all the parameters in the model. As the number of parameters increases, this leads to inefficiency and consumes significant resources[33].Therefore, we utilized Low-Rank Adaptation(LoRA)[34] to fine-tune the local open-source LLMs, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. LoRA can be applied to the $\boldsymbol{q}$, $\boldsymbol{k}$ and $\boldsymbol{v}$ matrices in attention calculation, which adds two structures, $\boldsymbol{A}$ and $\boldsymbol{B}$, next to the pre-trained model structure. For the linear layer $\boldsymbol{h} = \boldsymbol{W_0 x}$, the forward propagation is modified to: $\boldsymbol{h} = \boldsymbol{W_0 x} + \boldsymbol{BAx}$, where $\boldsymbol{W_0} \in \boldsymbol{R^{d\times k}}$, $\boldsymbol{B} \in \boldsymbol{R^{d\times r}}$, $\boldsymbol{A} \in \boldsymbol{R^{r\times k}}$, rank $\boldsymbol{r} \ll \mathbf{min}(\boldsymbol{d}, \boldsymbol{k})$.

During the instruction-tuning process, we fine-tuned LLMs on the mental health instruction datasets to obtain the unified IE model:

$$\boldsymbol{LLM_{ie}} = \boldsymbol{SFT}(\boldsymbol{LLM_{base}}, \boldsymbol{D_{ie}}) \tag{1}$$

where $\boldsymbol{LLM_{ie}}$ is the fine-tuned unified information extraction model, $\boldsymbol{LLM_{base}}$ is the pre-trained base model without fine-tuned, and $\boldsymbol{D_{ie}}$ is the mental health information extraction instruction datasets. We utilize LlamaFactory framework[35] to implement Instruction-tuning with LoRA.

3.2. Type-verification

The type-verification stage is employed to further improve the performance of $\boldsymbol{LLM_{ie}}$ in information extraction after instruction-tuning, which trains a classification model (e.g., BERT) with much smaller parameters than the $\boldsymbol{LLM_{ie}}$ used for instruction-tuning. The verification results are then returned to the $\boldsymbol{LLM_{ie}}$, further enhancing the performance of information extraction by answering again. We introduce this process in detail in this section. The data acquisition process of the type-verification component is shown in Fig. 3, which illustrates a simple example.

3.2.1. Sample construction and training

To obtain a training dataset for type-verification, a series of steps are required to construct it from the original instruction datasets. Initially, we extract task instructions $\boldsymbol{I_t}$ and their corresponding output instances $\boldsymbol{y}$ from the original instruction data. Subsequently, to construct a classification task dataset, we extract entity, relation, and event types from the output instances $\boldsymbol{y}$. When constructing positive samples, we modify the task instructions $\boldsymbol{I_t}$ to inquire whether they involve the extracted types, while simultaneously marking the output instances $\boldsymbol{y}$ as *"correct."* Conversely, for each positive sample constructed, we randomly select a sample with a different task instruction $\boldsymbol{I_t}$ to create negative samples. These samples do not match the originally correct type-verification task instructions, thereby annotating the type-verification results as *"incorrect"*.

Utilizing pre-trained models with small parameter sizes as auxiliary components can effectively enhance the performance and efficiency of LLMs [36]. Given this premise, we selected BERT as the base model for training the type-verification task due to its reasonably small 110M parameters size and exceptional performance in classification tasks [37]. To align with the data requirements of BERT, we concatenate the task instructions $\boldsymbol{I_t}$ and input instances $\boldsymbol{x}$ from the type-verification dataset, treating the output instances $\boldsymbol{y}$ as classification results, thus framing the task as a binary classification task: $\boldsymbol{Dtv} = (\boldsymbol{I_t} + \boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{Dtv}$ represents the classification dataset for type-verification. Following this, we partitioned $\boldsymbol{D_{tv}}$ into training, verification, and testing sets in a 6:2:2 ratio for fine-tuning the base BERT model. Ultimately, we obtained a lightweight model tailored for type-verification.

$$\boldsymbol{BERT_{tv}} = \boldsymbol{FT}(\boldsymbol{BERT_{base}}, \boldsymbol{D_{tv}}) \tag{2}$$

where $\boldsymbol{BERT_{base}}$ represents the pre-trained BERT model without fine-tuning, and $\boldsymbol{BERT_{tv}}$ is the refined BERT model specifically fine-tuned for type-verification datasets.

3.2.2. Application verification

The use of the type-verification component is shown in the Fig. 4. We first extract the type information for each sample from the results of $\boldsymbol{LLM_{ie}}$ information extraction. Subsequently, we subject them to type-verification using $\boldsymbol{BERT_{tv}}$:

$$\boldsymbol{y} = \boldsymbol{BERT_{tv}}(\boldsymbol{x}) \tag{3}$$

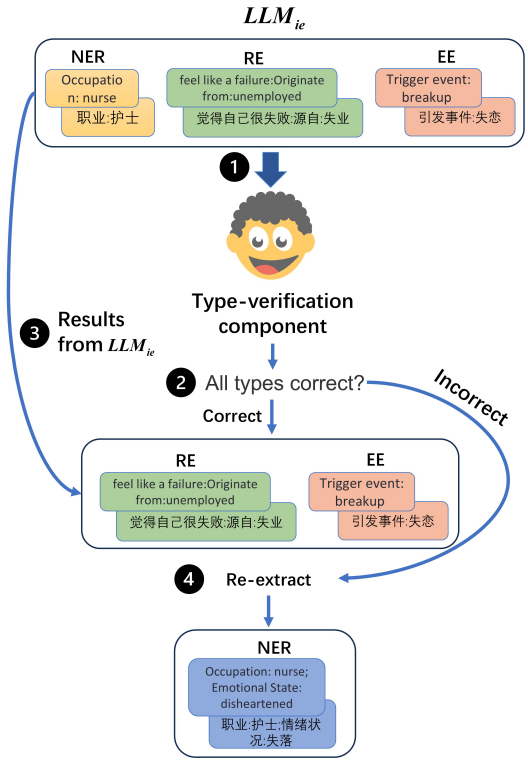

Fig. 4: type-verification component for implementing verification.

Where $\boldsymbol{x}$ represents the input and $\boldsymbol{y}$ denotes the output, i.e., the predicted result. If $\boldsymbol{BERTtv}$'s output $\boldsymbol{y}$ is "correct," we consider the type information extracted by $\boldsymbol{LLM_{ie}}$ to be accurate and directly output it as the final result. However, if $\boldsymbol{BERT_{tv}}$'s output $\boldsymbol{y}$ is "incorrect," it means that $\boldsymbol{LLM_{ie}}$ may have made an error in extracting the information. To correct this error, we use a technique called hard prompting to alert $\boldsymbol{LLM_{ie}}$ to the potential issues in type extraction and request it to re-extract the information. Specifically, we provide $\boldsymbol{LLM_{ie}}$ with explicit prompts indicating that the previously extracted type information may be incorrect and ask it to re-extract the information. The purpose of this method is to prompt the model to be more accurate in the re-extraction process by clearly pointing out the issues.

Through this approach, we can verify and correct $\boldsymbol{LLM_{ie}}$'s output in each information extraction task, significantly improving the performance of information extraction. Experimental results have shown that this method greatly enhances $\boldsymbol{LLM_{ie}}$'s performance in information extraction tasks, making the final extraction results more accurate and reliable. This process is shown in the lower part of Fig.1. Subsequent experiments have demonstrated a significant enhancement in $\boldsymbol{LLM_{ie}}$'s performance on information extraction tasks.

4. Experiments

In this section, we first introduce the foundation LLMs, evaluation metrics, and baseline methods used in our experiments. Next, we conduct experiments and analyses across three different information extraction tasks. We then examine the selection of TV models, perform ablation studies on TV, DA, and UIE, and conclude with a case study.

4.1. Experimental settings

4.1.1. Foundation LLMs

We selected Qwen-7B[1], Baichuan2-7B[2], ChatGLM3-6B[3], and Llama3-8B-Chinese-Chat[4] as foundation LLMs. Among these, Qwen-7B, developed by Alibaba Cloud, is a Transformer-based large language model pretrained on diverse data sources, including web texts, books, and code. Baichuan2-7B, developed by Baichuan Intelligence Inc., is trained on a high-quality corpus with 2.6 trillion tokens and has achieved state-of-the-art performance in authoritative Chinese and English benchmarks of its size. ChatGLM3-6B, the latest open-source model in the ChatGLM series, utilizes a more diverse training dataset, additional training steps, and an optimized training strategy, with evaluations spanning semantics, mathematics, reasoning, coding, and knowledge. Llama3-8B-Chinese-Chat is an instruction-tuned language model designed for both Chinese and English users, featuring capabilities such as roleplaying and tool usage, and is built upon the Meta-Llama-3-8B-Instruct model.

4.1.2. Evaluation metrics

To verify the accuracy of extraction, we apply $\boldsymbol{F_1}$ score as the evaluation metric for extraction accuracy.

$$\boldsymbol{P = \frac{T_p}{T_p + F_p} \times 100\%} \tag{4}$$

$$\boldsymbol{R = \frac{T_p}{T_p + F_N} \times 100\%} \tag{5}$$

$$\boldsymbol{F_1 = \frac{2P \times R}{P + R} \times 100\%} \tag{6}$$

Here, $\boldsymbol{P}$ and $\boldsymbol{R}$ represent precision and recall, respectively, where $\boldsymbol{T_P}$ denotes the number of correctly identified bi-gram or tri-gram tuples, $\boldsymbol{F_P}$ represents the number of incorrectly identified bi-gram or tri-gram tuples, and $\boldsymbol{F_N}$ represents the number of missed bi-gram or tri-gram tuples.

Furthermore, to validate the completeness of extraction, we also employ ROUGE-N as an evaluation metric.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \boldsymbol{Count_{match}(gram_n)}}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \boldsymbol{Count(gram_n)}} \tag{7}$$

Here, the denominator represents the total number of N-grams in the reference answers, while the numerator denotes

[1] https://huggingface.co/Qwen/Qwen-7B
[2] https://huggingface.co/baichuan-inc/Baichuan2-7B
[3] https://huggingface.co/THUDM/chatglm3-6b
[4] https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat

the count of N-grams shared between the predicted results generated by the model and the reference answers. In this paper, N is set to 2.

The overall extraction performance of the model is assessed by the weighted sum of these two metrics, resulting in a score denoted as Score.

$$\text{Score} = \alpha \cdot \boldsymbol{F_1} + \beta \cdot \text{ROUGE-N} \tag{8}$$

Here, α and β respectively represent the importance weights of the $\boldsymbol{F_1}$ score and ROUGE-N. In this paper, $\alpha = 0.5$ and $\beta = 0.5$, indicate that in the evaluation of this study, we place equal importance on extraction accuracy and completeness.

4.1.3. Baseline method

We employed two distinct approaches to compare and evaluate the performance of $\boldsymbol{LLM_{ie}}$ on the task of mental health information extraction. The first group involves information extraction methods based on auto-encoder pre-trained models.

BERT-BiLSTM-CRF[38]: The model utilizes a pre-trained Chinese BERT model to obtain character embeddings. The output vectors are then fed into a bidirectional long short-term memory (BiLSTM) network to extract the features required for entity recognition. Finally, the output is passed through a Conditional Random Field (CRF) layer for decoding and computing the optimal labeling sequence.

AdaSeq Bert-CRF[39]: Based on the Transformer-CRF model, this approach enhances model accuracy by incorporating relevant sentences recalled by external tools as additional context.

PRGC[40]: The joint relation triplet extraction model based on latent relationships and global correspondences.

OneRel[41]: The model that treats joint extraction as a fine-grained triplet classification problem, which greatly alleviates the problems of cascading errors and redundant information, achieving state-of-the-art performance for the WebNLG dataset.

DiffusionNER[42]: The model that formulates the named entity recognition task as a boundary-denoising diffusion process and thus generates named entities from noisy spans. It achieves comparable or even better performance than previous state-of-the-art models.

The second group involves a comparative evaluation based on different experimental settings of prediction.

Origin: The original base model without fine-tuning.

None-Ins.: No task-specific instruction descriptions provided during the fine-tuning process.

Uni-Ins.: Only one task-specific instruction description is provided for each task during the fine-tuning process.

4.1.4. Implementation details

All the models we utilized were downloaded from Huggingface, and all experiments were conducted on a single RTX4090 graphics card. For all LLMs employed in instruction-tuning, we maintained consistent hyperparameter selections. Some hyperparameter settings are listed in Table 3.

4.2. Main results

Table 2 shows the comparison results for all baseline methods on our datasets across various tasks. During the experiments, we observed that the model without any fine-tuning (Origin) exhibited almost no extraction capability and failed to produce outputs in the required format as per the instructions. Therefore, we only report its ROUGE scores.

Named Entity Recognition (NER): Among the three tasks, the model performs best on the NER task, which we attribute to the relatively clear boundaries of most extracted types in this task; for instance, emotional states are typically demarcated by terms like *'sad,' 'grieving,'* etc. This suggests that the $\boldsymbol{LLM_{ie}}$ can swiftly identify personal information within the mental health domain.

From our experimental results presented in Table 2, it is evident that instruction-tuning the Baichuan2-7B model yielded remarkable performance, attaining a Score of 88.40, which is the highest among all models tested. This performance surpasses auto-encoder models such as BiLSTM-CRF, AdaSeq CRF, and DiffusionNER. The instruction-tuned Baichuan2-7B model not only achieved the highest Score but also exhibited superior performance in terms of $\boldsymbol{F_1}$ metrics. All instruction-tuned models, except ChatGLM3-6B and Llama3-8B-Chinese-Chat, outperformed their auto-encoder counterparts.

The performance of the models significantly declined under the None-Ins. and Uni-Ins. methods. The Baichuan2-7B model exhibited notable decreases in Score, $\boldsymbol{F_1}$, and ROUGE-2. The Qwen-7B model demonstrated moderate reductions across these metrics. ChatGLM3-6B recorded the most pronounced drops in all measures. Finally, Llama3-8B-Chinese-Chat experienced reductions as well, albeit to a lesser extent.

Relation Extraction (RE): The RE task proves most challenging, as it entails extracting triples that not only require identifying subjects and objects but also discerning the relation between them. Complicating matters, both subjects and objects often span extensive and ambiguous textual segments. For example, both *'experiencing the failure of the exam'* and *'the failure of the exam'* can serve as the same triggering event for the current emotion. This ambiguity in span poses a considerable obstacle in the RE task. Consequently, its scores are lower compared to the other three tasks.

In our experiments, the instruction-tuned Llama3-8B-Chinese-Chat model achieved a score of 70.78, marking the highest level of performance among the models evaluated. It also achieved $\boldsymbol{F_1}$ and ROUGE-2 of 82.21 and 70.78, respectively. This superior performance surpasses the auto-encoder models, which achieve state-of-the-art performance in other RE datasets. Moreover, all instruction-tuned models showed enhanced performance compared to auto-encoder models.

Despite this, the None-Ins. and Uni-Ins. methods led to varying degrees of performance degradation across models. Baichuan2-7B and Qwen-7B exhibited moderate declines in Score, $\boldsymbol{F_1}$, and ROUGE-2, while ChatGLM3-6B experienced the most substantial performance drop. Llama3-8B-Chinese-Chat also showed reductions, though they were less pronounced compared to ChatGLM3-6B.

Table 2: Results of mental health information extraction task. The bold represents the best performance of all baselines.

	Method	NER			RE			EE		
		F_1	ROUGE-2	Score	F_1	ROUGE-2	Score	F_1	ROUGE-2	Score
Auto-encoder models	BiLSTM-CRF	87.70	85.64	86.67	47.34	71.58	59.46	68.34	85.88	77.11
	AdaSeq CRF	87.29	**89.18**	88.24	-	-	-	73.25	86.88	80.07
	PRGC	-	-	-	56.79	75.87	66.33	-	-	-
	OneRel	-	-	-	55.81	74.56	65.19	-	-	-
	DiffusionNER	86.93	86.42	86.68	-	-	-	77.39	86.72	82.06
Baichuan2-7B	Origin	-	9.02	-	-	18.31	-	-	25.68	-
	None-Ins.	77.43	78.80	78.11	44.29	69.49	56.89	65.82	81.95	78.89
	Uni-Ins.	83.64	83.07	83.36	52.97	78.96	65.97	75.64	87.60	81.62
	Ours	**87.75**	89.04	**88.40**	56.13	80.01	68.07	76.59	88.29	82.44
Qwen-7B	Origin	-	26.94	-	-	15.41	-	-	31.45	-
	None-Ins.	83.07	83.68	83.38	50.90	74.52	62.71	71.85	84.73	78.29
	Uni-Ins.	85.26	84.98	85.12	53.07	78.07	65.57	73.70	85.21	79.46
	Ours	87.43	89.05	88.24	54.80	80.88	67.84	75.26	86.70	80.98
ChatGLM3-6B	Origin	-	19.02	-	-	20.66	-	-	32.80	-
	None-Ins.	56.67	63.98	60.33	19.49	50.42	34.96	55.51	78.47	66.94
	Uni-Ins.	81.56	81.29	81.43	50.62	76.98	63.80	72.74	86.25	79.50
	Ours	86.69	88.21	87.45	58.78	81.12	69.95	77.00	88.64	82.82
Llama3-8B-Chinese-Chat	Origin	-	16.28	-	-	16.89	-	-	19.24	-
	None-Ins.	78.43	79.27	78.85	53.56	75.61	64.59	73.93	86.38	80.16
	Uni-Ins.	80.26	82.90	81.58	86.33	79.27	67.80	77.84	88.26	83.50
	Ours	86.37	88.02	87.20	**58.74**	**82.21**	**70.78**	**78.95**	**88.21**	**83.58**

Table 3: Hyperparameter setting.

LLM_{ie}		$BERT_{tv}$	
Learning rate	1e-4	Learning rate	3e-5
Batch size	8	Batch size	16
Epochs	15	Epochs	15
Lora rank	8		

Event Extraction (EE): Under the EE task, only the event type and its corresponding event need to be extracted. However, compared to the NER task, the spans are generally longer and the boundaries more ambiguous. For instance, both *'recent fractures in a relationship'* and *'fractures in a relationship'* can qualify as trigger events for physical or emotional conditions. Consequently, it achieves a middle-ranking score among the three tasks.

In the EE task, the instruction-tuned Llama3-8B-Chinese-Chat model achieved a Score of 83.58, with an F_1 of 88.21 and a ROUGE-2 of 83.58. This result is the highest among all the models tested, outperforming auto-encoder models. Other instruction-tuned models also performed close to the DiffusionNER model, which achieves comparable or even better performance than previous state-of-the-art models. These results indicate their competitive edge.

Nonetheless, the None-Ins. and Uni-Ins. methods resulted in varying degrees of performance decline across all evaluated models. Specifically, Baichuan2-7B, Qwen-7B, and Llama3-8B-Chinese-Chat exhibited moderate drops in Score, F_1, and ROUGE-2, while ChatGLM3-6B experienced a pronounced decline across all metrics.

Overall, the experimental results clearly demonstrate that instruction-tuning significantly enhances the performance of models across NER, RE, and EE tasks, outperforming traditional auto-encoder models. However, the None-Ins. and Uni-Ins. methods lead to noticeable performance declines. We analyzed the None-Ins. method does not provide instructions in the process of instruction-tuning, which results in LLMs learning solely based on data features, subsequently hindering their ability to generate answers in response to user instructions. In contrast, the Uni-Ins. method, which offers only a single instruction during instruction-tuning, diminishes LLMs' instruction-following capability and instruction generalization, ultimately leading to a decline in information extraction performance. The above analysis shows the importance of instruction-tuning with multi-instructions for optimal results.

4.3. Analysis and discussion

4.3.1. Selection of type-verification model

We previously selected the pre-trained model BERT as the classifier for the type-verification task. To comprehensively validate this choice, we selected four models with different parameters and architectures to study: specifically, BERT (110M, Encoder-only), T5 (770M, Encoder-Decoder), BLOOMZ-3B (Decoder-only), and Baichuan2-7B (Decoder-only). The experimental parameters of the above models follow Table 3. Then,

Table 4: Ablation study for the components of type-verification (TV).

		NER			RE			EE		
		F_1	ROUGE-2	Score	F_1	ROUGE-2	Score	F_1	ROUGE-2	Score
Baichuan2-7B	w/o TV	86.86	88.10	87.48	54.26	78.96	66.61	76.26	86.88	81.57
	w/ TV	**87.75**	**89.04**	**88.40**	**56.16**	**80.01**	**68.07**	**76.59**	**88.29**	**82.44**
Qwen-7B	w/o TV	87.08	88.22	87.65	54.01	78.33	66.17	74.30	85.26	79.78
	w/ TV	**87.43**	**89.05**	**88.24**	**54.80**	**80.88**	**67.84**	**75.26**	**86.70**	**80.98**
ChatGLM3-6B	w/o TV	85.33	86.84	86.09	56.01	79.89	67.95	75.15	88.41	81.65
	w/ TV	**86.69**	**88.21**	**87.45**	**58.78**	**81.12**	**69.95**	**77.00**	**88.64**	**82.82**
Llama3-8B-Chinese-Chat	w/o TV	86.31	87.65	86.98	58.03	81.08	69.56	78.83	87.03	82.93
	w/ TV	**86.37**	**88.02**	**87.20**	**58.74**	**82.21**	**70.78**	**78.95**	**88.21**	**83.58**

Table 5: Ablation study for data augmentation (DA).

		NER			RE			EE		
		F_1	ROUGE-2	Score	F_1	ROUGE-2	Score	F_1	ROUGE-2	Score
Baichuan2-7B	w/o DA	85.05	87.14	86.10	44.28	77.33	60.81	54.32	83.81	69.07
	w/ DA	**85.91**	**88.25**	**87.08**	**49.11**	**78.69**	**63.90**	**62.63**	**87.74**	**75.19**
Qwen-7B	w/o DA	84.85	87.63	86.24	45.52	77.08	61.30	54.33	84.29	69.31
	w/ DA	**85.60**	**88.25**	**86.93**	**52.11**	**79.30**	**65.71**	**62.76**	**86.49**	**74.63**
ChatGLM3-6B	w/o DA	84.33	87.45	85.89	43.60	75.82	59.71	52.31	83.44	67.88
	w/ DA	**85.63**	**87.77**	**86.70**	**47.25**	**78.52**	**62.89**	**59.73**	**84.43**	**72.08**
Llama3-8B-Chinese-Chat	w/o DA	83.94	86.80	85.37	46.63	**78.34**	62.49	55.04	84.02	69.53
	w/ DA	**85.88**	**88.39**	**87.14**	**51.73**	78.02	**64.88**	**61.55**	**86.28**	**73.92**

Table 6: Ablation study for unified information extraction (UIE).

		NER			RE			EE		
		F_1	ROUGE-2	Score	F_1	ROUGE-2	Score	F_1	ROUGE-2	Score
Baichuan2-7B	w/o NER	-	5.81	-	53.21	78.41	65.81	76.10	86.53	81.32
	w/o RE	83.24	86.03	84.64	-	35.01	-	73.13	84.38	78.76
	w/o EE	85.51	88.34	86.93	51.31	75.99	63.65	-	20.67	-
	w/ UIE	**87.75**	**89.04**	**88.40**	**56.16**	**80.01**	**68.07**	**76.59**	**88.29**	**82.44**
Qwen-7B	w/o NER	-	6.94	-	**54.91**	78.32	66.62	74.36	86.48	80.42
	w/o RE	83.34	84.20	83.77	-	23.67	-	71.06	82.68	77.37
	w/o EE	82.46	83.15	82.81	54.26	78.96	66.61	-	27.29	-
	w/ UIE	**87.43**	**89.05**	**88.24**	54.80	**80.88**	**67.84**	**75.26**	**86.70**	**80.98**
ChatGLM3-6B	w/o NER	-	10.26	-	53.56	78.44	66.00	76.35	87.06	81.71
	w/o RE	86.76	**88.45**	**87.61**	-	30.24	-	73.16	85.78	79.32
	w/o EE	86.36	87.83	87.10	56.26	80.98	68.62	-	26.98	-
	w/ UIE	**86.69**	88.21	87.45	**58.78**	**81.12**	**69.95**	**77.00**	**88.64**	**82.82**
Llama3-8B-Chinese-Chat	w/o NER	-	9.34	-	56.46	80.96	68.71	76.56	87.03	81.80
	w/o RE	86.86	88.10	87.48	-	35.37	-	73.28	83.78	78.83
	w/o EE	**86.86**	87.45	87.16	56.66	81.94	69.30	-	29.67	-
	w/ UIE	86.37	**88.02**	**87.20**	**58.74**	**82.21**	**70.78**	**78.95**	**88.21**	**83.58**

we present our arguments from two perspectives: **Classification accuracy** and **computational cost**.

Classification accuracy: we fine-tuned models with varying parameter counts and conducted performance comparisons on the type-verification datasets we constructed. As shown in Table 7, BERT exhibits the highest classification accuracy in this task. In comparison to text generation models, BERT demonstrates superior performance as a classifier.

Computational cost: we documented the GPU resources consumed and the training durations during the fine-tuning process of each model. As illustrated in Table 7, BERT displays a notable advantage in both GPU resource consumption and training duration. These findings indicate that employing BERT as the classifier for the type validation task not only enhances performance but also improves efficiency.

Table 7: Classification accuracy and hardware resource consumption for different models in type-verification task.

	BERT	T5	BLOOMZ-3B	Baichuan2-7B
Classification accuracy	88%	86%	81%	83%
Training times	16m	28m	1h28m16s	1h58m37s
GPU memory	7G	11.9G	17.3G	19.5G
Parameters	110M	770M	3B	7B

4.3.2. Ablation study

We first perform ablation experiments on the component of the type-verification to study its validity and finally obtain the results on our mental health information extraction datasets, as shown in Table 4. The addition of the type-verification component contributed significantly to the performance improvements across all $\boldsymbol{LLM_{ie}}$ models. We consider that type-verification ensures the consistency and correctness of the extracted entity, relation, and event types, thereby reducing errors and noise. Furthermore, type-verification acts as a filter, helping the models to eliminate incorrect information that does not meet the expected types, thus increasing the precision of information extraction. Therefore, when the type-verification component is removed, all $\boldsymbol{LLM_{ie}}$ models exhibit a significant decline in performance across NER, RE, and EE tasks. This demonstrates the benefits of the type-verification component.

To validate the efficacy of the data augmentation strategies proposed in this paper, we also conducted an ablation study on data augmentation. Specifically, we randomly selected 100 samples from the test set and subjected them to operations including synonym replacement, random insertion, random exchange, and random deletion, simulating certain text distortions encountered in real life. Subsequently, ablation experiments were conducted on each of the four LLMs respectively. As shown in Table 5: after being fine-tuned with data augmentation, all models were capable of learning contextual and semantic information from perturbed data, thereby becoming more robust. A notable improvement in performance relative to models fine-tuned without data augmentation was observed, which substantiates the effectiveness of our data augmentation strategy.

To validate the necessity of unified information extraction, we selectively disabled one task at a time during the instruction-tuning phase and monitored the performance changes. The results are depicted in Table 6. In the process of unified information extraction, there exists interdependency and interaction among different tasks, where each task benefits from supplementary and supportive information from others. For instance, in the RE task, the identification of the relation between subjects and objects facilitated by the NER task enables a more precise determination of entity boundaries and types. Similarly, in the EE task, entities participating in events provided by the NER task further aid in specifying event types and details. When a task is omitted, LLMs fail to capitalize fully on the information supplied by other tasks. Consequently, the removal of a single task during instruction-tuning not only degrades the performance of LLMs on the test set of the removed task but also adversely affects the performance on other tasks, underscoring the importance of unified information extraction.

4.3.3. Confidence Interval Analysis

To further validate the stability and reliability of our experimental results, we performed a confidence interval (CI) analysis on the $\boldsymbol{F_1}$ and ROUGE-2 scores. We repeated each experiment five times with different random seeds, recorded the performance distribution of some combinations of models and tasks, and calculated the 95% CI based on the mean and standard error. The steps are as follows:

- We performed five independent runs with different random seeds and recorded the $\boldsymbol{F_1}$ and ROUGE-2 scores for each run on Baichuan2-7B on the NER task, Qwen-7B on the RE task, and Llama3-8B-Chinese-Chat on the EE task.
- Then we computed the mean($\boldsymbol{\mu}$) and standard error ($\boldsymbol{SE = \frac{\sigma}{\sqrt{n}}}$, where $\boldsymbol{\sigma}$ is the standard deviation and $\boldsymbol{n = 5}$).The 95% CI was calculated using the formula: $\boldsymbol{CI = \mu \pm 1.96 \times SE}$.

Table 8 summarizes the 95% confidence intervals for $\boldsymbol{F_1}$ and ROUGE-2 scores across three models. The fluctuation ranges of the CIs are consistently within 1%, demonstrating minimal variability and high stability in the experimental results.

Table 8: Confidence Intervals for $\boldsymbol{F_1}$ and ROUGE-2 scores.

Model	$\boldsymbol{F_1}$(95% CI)	ROUGE-2(95% CI)
Baichuan2-7B(NER)	87.80 ± 0.23	88.95 ± 0.52
Qwen-7B(RE)	54.82 ± 0.56	80.69 ± 0.71
Llama3-8B-Chinese-Chat(EE)	79.04 ± 0.39	88.15 ± 0.65

4.4. Case study

To illustrate more intuitively the changes in extraction results of $\boldsymbol{LLM_{ie}}$ before and after type verification, we conducted a case study on three sentences, as shown in Table 9. In the first and second sentences, After incorporating the type-verification component, $\boldsymbol{LLM_{ie}}$ transformed previously incorrect outputs into correct results, demonstrating the effectiveness of this component. In the third sentence, although the type verification

Table 9: Case study on five sentences.

Sentence	Type-verification	w/o Type-verification	Type-verification
Recently, I have been feeling very anxious and uneasy, and I am unsure how to cope with it. As a 24-year-old university student, I am facing academic and interpersonal pressures. I often worry about my performance and future prospects, and this anxiety has affected my studies and daily life. I hope to find some methods to alleviate my anxiety and restore my inner peace. 我最近感到很焦虑和不安，不知道该如何应对。我是一个24岁的大学生，面临着学业和人际关系的压力。我常常担心自己的表现和未来的发展，这种焦虑情绪影响了我的学习和生活。我希望能够找到一些方法来缓解焦虑，恢复内心的平静。 *Lable:{Age:24;Occupation:University Student;Emotional state:Anxious}* *Lable:*{年龄*:24*岁*;*职业*:*大学生*;*情绪状况*:*焦虑*}*	Incorrect	{*Age:24;Emotional state:anxious;Emotional state:Uneasy*} {年 龄:24岁;情 绪 状 况:焦虑;情绪状况:不安}	{*Age:24;Occupation:University Student;Emotional state:Anxious*} {年 龄:24岁;职 业:大 学生;情绪状况:焦虑}
I am a 33-year-old actor, and I have been feeling very anxious and stressed recently. The demands of stage performances and public scrutiny make me feel nervous and uneasy. These pressures have led to insomnia and a loss of appetite, affecting both my acting skills and mental state. 我是一名33岁的演员，最近感到很焦虑和压力大。工作上需要面对舞台表演和公众的评判，让我感到很紧张和不安。这些压力让我失眠和食欲不振，影响了我的演技和心理状态。 *Lable:{Age:33;Occupation:Actor;Physical condition:Insomnia;Physical condition:Loss of Appetite;Emotional state:Nervous}* *Lable:*{年 龄*:33*岁*;*职 业*:*演 员*;*躯 体 状 况*:*失 眠*;*躯 体 状 况*:*食 欲 不振*;*情绪状况*:*焦虑*}*	Incorrect	{*Age:33;Occupation:Actor; Emotional state:Nervous*} {年 龄:33岁;职 业:演 员;情绪状况:焦虑}	{*Age:33;Occupation:Actor; Physical condition:Insomnia;Physical condition:Loss of Appetite;Emotional state:Nervous*} {年 龄:33岁;职 业:演 员;躯体状况:失眠;躯体状况:食欲不振;情绪状况:焦虑}
I am concerned and uneasy about my physical condition. As a 34-year-old doctor, long working hours have led to feelings of fatigue and discomfort. I hope to find some methods to improve my physical condition and maintain health and vitality. 我对自己的身体状况感到担忧和不安。我是一个34岁的医生，长时间的工作导致我感到身体疲劳和不适。我希望能够找到一些方法来改善身体状况，保持健康和活力。 *Lable:{Age:34;Occupation:Doctor;Physical condition:Fatigue}* *Lable:*{年龄*:34*岁*;*职业*:*医生*;*躯体状况*:*身体疲劳*}*	Incorrect	{*Age:34;Occupation:Doctor; Physical condition:Fatigue*} {年 龄:34岁;职 业:医 生;躯体状况:身体疲劳}	{*Age:34;Occupation:Doctor; Physical condition:Fatigue*} {年 龄:34岁;职 业:医 生;躯体状况:身体疲劳}
I am a physician who is constantly exposed to the pain of patients and suffering families. These situations leave me feeling frustrated and helpless. 我是一名医生，经常接触到病人的痛苦和痛苦的家庭。这些情况让我感到沮丧和无助。 *Lable:{Occupation:Doctor;Emotional state:Frustrated}* *Lable:*{职业*:*医生*;*情绪状况*:*沮丧*}*	Correct	{*Occupation:Doctor;Emotional state:Frustrate;Emotional state:helpless*} {职 业:医 生;情 绪 状 况:沮丧;情绪状况:无助}	{*Occupation:Doctor;Emotional state:Frustrate;Emotional state:helpless*} {职 业:医 生;情 绪 状 况:沮丧;情绪状况:无助}
I was late this morning and was criticized by my boss. I began to doubt my abilities and felt suspended for being useless. This doubt created a very frustrated mood in me. 今天早上迟到了，被老板批评了一顿。我开始怀疑自己的能力，觉得暂停自己一无是处。这种怀疑让我产生了很沮丧的情绪。 *Lable:{Criticized by my boss:Originate from:Late;Late:Produce:Frustrated}* *Lable:*{被老板批评*:*源自*:*迟到*,*迟到*:*产生*:*沮丧*}*	Incorrect	{*Late:Produce:Frustrated*} {迟到:产生:沮丧}	{*Late:Produce:Frustrated*} {迟到:产生:沮丧}

component classifies the outputs as *"Incorrect,"* $\boldsymbol{LLM_{ie}}$'s output remains unchanged and still matches the labels. Our analysis suggests that $\boldsymbol{LLM_{ie}}$ is reluctant to alter answers with high confidence, indicating that even when type-verification fails, $\boldsymbol{LLM_{ie}}$ still can produce correct results.

The fourth and fifth sentences are examples of extraction errors, the former is predicted as *"correct"* by the TV component resulting in no further correction by LLMs, while for the latter, even if the TV component judges it as *"incorrect,"* the LLMs do not correct its answer in further extraction.

4.5. Discussion and future work for TV component

In our experiments, we have identified certain limitations of the TV component. Firstly, the TV component demonstrates limited performance improvement in some cases. We have determined that there are two primary factors contributing to this situation. One is that the component is designed solely for determining the correctness of the type. When the type is correct but the quantity is incorrect, it is unable to perform further corrections. Secondly, the limited generalization capability of LLMs may result in failure to rectify errors. For instance, sentences exhibiting substantial divergence from training data patterns often evade effective correction by LLMs, even after TV component verification.

Additionally, the information extraction task using the TV component requires one prediction and two extractions, its integration inevitably increases the overall processing time of the workflow. This added time may pose challenges for time-sensitive applications, particularly those requiring real-time or near-real-time responses. Scaling the method to handle larger datasets or deploying it in complex environments could further exacerbate these time constraints, necessitating additional computational optimizations. Furthermore, the method has not yet been tested on datasets from other domains, leaving its adaptability and effectiveness in diverse fields unexplored.

In future, we consider developing a dynamic verification framework that combines type-verification with quantitative error detection. And validate the method's performance on real-world datasets across various domains and explore strategies to improve the balance between computational efficiency and scalability, ensuring its practicality in high-stakes scenarios.

5. Conclusion

This study presents an innovative solution to enhance the performance of LLMs in extracting information from unstructured mental health texts. By constructing a synthetic Chinese mental health dataset under the guidance of psychology experts, we addressed both data scarcity and privacy concerns in the Chinese mental health domain. Our approach introduces a lightweight type-verification component that improves extraction performance while reducing computational requirements. Experimental results demonstrate that the type-verification component consistently enhances the accuracy of various LLMs in mental health information extraction. This modular design not only improves the reliability of LLM outputs but also demonstrates how smaller, domain-specific models can complement larger LLMs to achieve greater efficiency.

The findings have broad implications beyond the mental health domain, with potential applications in high-stakes fields such as law, finance, and education. For instance, the proposed approach could enhance precision in legal document analysis, improve financial data extraction, and support personalized education systems. This study highlights how domain-specific models, when combined with interdisciplinary collaboration, can enable reliable AI solutions for addressing complex real-world challenges.

Limitations

While the proposed method demonstrates strong performance in mental health information extraction, several limitations should be considered. The dataset used for instruction-tuning was generated by LLMs for a limited set of initial instances. Although our work safeguards privacy and addresses the scarcity of Chinese instruction datasets, the generated data may not fully capture the diversity and complexity of real-world counseling texts, potentially limiting the model's generalizability across various scenarios or edge cases. Furthermore, as our data construction method is focused on the mental health domain, experiments in other fields have not yet been conducted, which may present challenges for broader applications.

Ethics

The use of AI in mental health research demands careful consideration of ethical principles, particularly regarding data privacy, informed consent, and fairness. In this study, all initial examples used to construct the instruction dataset were collected with explicit informed consent from participants, adhering to privacy regulations and ensuring the confidentiality and security of sensitive mental health information.

We recognize the inherent limitations of AI models, including potential biases in data and decision-making processes. While every effort was made to enhance transparency and mitigate bias, AI should complement—not replace—human judgment. The models are designed to support the expertise of mental health professionals, with results requiring interpretation by qualified professionals to provide appropriate context and informed decisions.

Finally, we underscore the need for strong ethical oversight, ensuring all AI applications in mental health pass ethical reviews to protect participants and uphold research integrity.

Data availability

We released the constructed mental health instruction dataset, instructions and related codes on https://github.com/CCzzzzzzz/mental_health.

Funding

This work was supported by National Science and Technology Major Project (2022ZD0116308); Minjiang University Science and Technology Pre-research Initiative for Recruited Talents (MJY21032, MJY23033); Fuzhou Marine Research Institute "Challenge-Driven and Talent-Led" Initiative (2024F02); Fujian Education Science Research Program for Young and Middle-Aged Teachers (JAT231095).

Appendix A. Data statistics

Table A.1: Types and quantity statistics of three information extraction tasks.

Task	Types	Numbers
NER	Occupation	674
	Physical condition	867
	Emotional state	986
	Family structure	155
	Thoughts	134
RE	Family closeness	25
	Family distance	35
	Peer closeness	16
	Peer distance	90
	Originate from	686
	Produce	778
EE	Trigger event	711
	Past experience	645

The types and quantity statistics of the various information extraction tasks are presented in Table A.1.

References

[1] A. D. Bergin, E. P. Vallejos, E. B. Davies, D. Daley, T. Ford, G. Harold, S. Hetrick, M. Kidner, Y. Long, S. Merry, et al., Preventive digital mental health interventions for children and young people: a review of the design and reporting of research, NPJ digital medicine 3 (2020) 133.

[2] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, D. Wang, Mental-llm: Leveraging large language models for mental health prediction via online text data, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8 (2024) 1–32.

[3] J. Cabrera, M. S. Loyola, I. Magaña, R. Rojas, Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots, in: International Work-Conference on Bioinformatics and Biomedical Engineering, Springer, 2023, pp. 313–326.

[4] H. Sun, Z. Lin, C. Zheng, S. Liu, M. Huang, PsyQA: A Chinese dataset for generating long counseling text for mental health support, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 1489–1503. URL: https://aclanthology.org/2021.findings-acl.130. doi:10.18653/v1/2021.findings-acl.130.

[5] H. Qiu, H. He, S. Zhang, A. Li, Z. Lan, Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support, arXiv preprint arXiv:2305.00450 (2023).

[6] Z. Cai, H. Fang, J. Liu, G. Xu, Y. Long, Unified information extraction in mental health domain based instruction tuning of large language models, Journal of Chinese Information Processing 38 (2024) 112–127.

[7] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Transactions on Knowledge Discovery from Data (2023).

[8] J. Li, J. Zhu, Q. Bi, G. Cai, L. Shang, Z. Dong, X. Jiang, Q. Liu, MINER: Multi-interest matching network for news recommendation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 343–352. URL: https://aclanthology.org/2022.findings-acl.29. doi:10.18653/v1/2022.findings-acl.29.

[9] H. Yan, J. Dai, T. Ji, X. Qiu, Z. Zhang, A unified generative framework for aspect-based sentiment analysis, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 2416–2429. URL: https://aclanthology.org/2021.acl-long.188. doi:10.18653/v1/2021.acl-long.188.

[10] P.-L. H. Cabot, R. Navigli, Rebel: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2370–2381.

[11] H. Fang, G. Xu, Y. Long, Y. Guan, X. Yang, Z. Chen, A system review on bootstrapping information extraction, Multimedia Tools and Applications 83 (2024) 38329–38353.

[12] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, H. Wu, Unified structure generation for universal information extraction, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5755–5772. URL: https://aclanthology.org/2022.acl-long.395. doi:10.18653/v1/2022.acl-long.395.

[13] J. Lou, Y. Lu, D. Dai, W. Jia, H. Lin, X. Han, L. Sun, H. Wu, Universal information extraction as unified semantic matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13318–13326.

[14] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, et al., Instructuie: multi-task instruction tuning for unified information extraction, arXiv preprint arXiv:2304.08085 (2023).

[15] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al., Instruction tuning for large language models: A survey, arXiv preprint arXiv:2308.10792 (2023).

[16] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13484–13508. URL: https://aclanthology.org/2023.acl-long.754. doi:10.18653/v1/2023.acl-long.754.

[17] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al., Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

[18] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, in: International Conference on Machine Learning, PMLR, 2023, pp. 22631–22648.

[19] L. Wang, R. Li, Y. Yan, Y. Yan, S. Wang, W. Wu, W. Xu, Instructionner: A multi-task instruction-based generative framework for few-shot ner, arXiv preprint arXiv:2203.03903 (2022).

[20] X. Li, K. Chen, Y. Long, M. Zhang, Llm with relation classifier for document-level relation extraction, arXiv preprint arXiv:2408.13889 (2024).

[21] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, Gpt-re: In-context learning for relation extraction using large language models, arXiv preprint arXiv:2305.02105 (2023).

[22] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 3534–3547.

[23] D. Hu, B. Liu, X. Zhu, X. Lu, N. Wu, Zero-shot information extraction from radiological reports using chatgpt, International Journal of Medical Informatics 183 (2024) 105321.
[24] O. Rohanian, M. Nouriborji, S. Kouchaki, F. Nooralahzadeh, L. Clifton, D. A. Clifton, Exploring the effectiveness of instruction tuning in biomedical language processing, Artificial intelligence in medicine (2024) 103007.
[25] S. Datta, K. Lee, H. Paek, F. J. Manion, N. Ofoegbu, J. Du, Y. Li, L.-C. Huang, J. Wang, B. Lin, et al., Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models, Journal of the American Medical Informatics Association 31 (2024) 375–385.
[26] M. Zhang, G. Jiang, S. Liu, J. Chen, M. Zhang, Llm–assisted data augmentation for chinese dialogue–level dependency parsing, Computational Linguistics (2024) 1–24.
[27] J. Yuan, R. Tang, X. Jiang, X. Hu, Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability, in: American Medical Informatics Association (AMIA) Annual Symposium, 2023.
[28] E. Strelcenia, S. Prakoonwit, Improving classification performance in credit card fraud detection by using new data augmentation, AI 4 (2023) 172–198.
[29] R. Tang, X. Han, X. Jiang, X. Hu, Does synthetic data generation of llms help clinical text mining?, arXiv preprint arXiv:2303.04360 (2023).
[30] J. Al-Garaady, M. Mahyoob, Chatgpt's capabilities in spotting and analyzing writing errors experienced by efl learners, Arab World English Journals, Special Issue on CALL (2023).
[31] K. Li, Y. Hu, L. He, F. Meng, J. Zhou, C-LLM: Learn to check Chinese spelling errors character by character, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 5944–5957. URL: https://aclanthology.org/2024.emnlp-main.340. doi:10.18653/v1/2024.emnlp-main.340.
[32] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.
[33] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6268–6278. URL: https://aclanthology.org/2023.emnlp-main.385. doi:10.18653/v1/2023.emnlp-main.385.
[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[35] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, Y. Ma, 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), Association for Computational Linguistics, Bangkok, Thailand, 2024. URL: http://arxiv.org/abs/2403.13372.
[36] B. Tan, Y. Zhu, L. Liu, E. Xing, Z. Hu, J. Chen, Cappy: Outperforming and boosting large multi-task lms with a small scorer, Advances in Neural Information Processing Systems 36 (2024).
[37] E. C. Garrido-Merchan, R. Gozalo-Brizuela, S. Gonzalez-Carvajal, Comparing bert against traditional machine learning models in text classification, Journal of Computational and Cognitive Engineering 2 (2023) 352–356.
[38] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).
[39] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, K. Tu, Improving named entity recognition by external context retrieving and cooperative learning, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 1800–1812. URL: https://aclanthology.org/2021.acl-long.142. doi:10.18653/v1/2021.acl-long.142.
[40] H. Zheng, R. Wen, X. Chen, Y. Yang, Y. Zhang, Z. Zhang, N. Zhang, B. Qin, X. Ming, Y. Zheng, PRGC: Potential relation and global correspondence based joint relational triple extraction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 6225–6235. URL: https://aclanthology.org/2021.acl-long.486. doi:10.18653/v1/2021.acl-long.486.
[41] Y.-M. Shang, H. Huang, X. Mao, Onerel: Joint entity and relation extraction with one module in one step, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 11285–11293.
[42] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, Diffusionner: Boundary diffusion for named entity recognition, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 3875–3890.