Research on Structured Item Information Extraction from Clinical Scale Texts Using ChatGPT and Zero-Shot Prompting

Authors: Hao Jie, Mo Zhiqiang, Sun Haixia, Chen Zhenli, Li Jiao, Sun Haixia

Date: 2024-08-06T00:00:00+00:00

Abstract

[Purpose/Significance] To achieve the extraction of structured item information from free clinical scale texts using ChatGPT without annotated data, thereby efficiently facilitating the structuring and intellectualization of medical scale resources.

[Methods/Process] We defined an item information extraction framework comprising 8 attribute types that accounts for structural differences in measurement concepts across clinical scales and collected documents from 59 clinically commonly used psychological assessment scales to construct a self-built dataset; designed categorized zero-shot prompts and conducted experiments using the official APIs of ChatGPT-3.5 and ChatGPT-4; and conducted a multi-angle analysis of the extraction performance of different ChatGPT versions when processing various clinical scale texts and possible influencing factors.

[Results/Conclusion] The source attribute extraction achieved the best performance, with Micro-F1 and Macro-F1 reaching at minimum 98.90% and 97.83%, respectively; followed by response options, usage instructions, and scoring rules; item numbers and item instructions ranked in the middle; clinical interpretation was the lowest, with Micro-F1 and Macro-F1 at 47.73% and 45.51%, respectively. ChatGPT-4 demonstrated overall superior performance, though the recall rate for some attributes was lower than that of ChatGPT-3.5. Increases in the hierarchical levels of scale measurement concepts, number of dimensions, number of items, and text length would degrade model performance. In summary, ChatGPT can efficiently assist in the structuring of medical scale resources, particularly when processing simple scales.

Full Text

Preamble

ChatGPT and Zero-Shot Prompt-based Structured Information Extraction for Clinical Scale Items

Jie Hao[1], Zhiqiang Mo[2], Haixia Sun[1], Zhenli Chen[1], Jiao Li[1]

[1]Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, 100020
[2]Department of Computer Science, University of Science and Technology of China, Hefei, 230026

Abstract

[**Purpose/Significance**] This study aims to extract structured item information from free-text clinical scales using ChatGPT without annotations, efficiently advancing the structuring and intellectualization of medical scale resources. [**Method/Process**] We defined an item information extraction framework comprising eight attribute types that accommodates structural differences in clinical scale measurement concepts. A dataset was constructed by collecting 59 commonly used clinical psychometric assessment scale documents. Zero-shot prompts were designed based on measurement concept levels, and experiments were conducted using the official ChatGPT-3.5 and ChatGPT-4 interfaces. We analyzed the extraction performance of different ChatGPT versions when processing various clinical scale texts and identified possible influencing factors from multiple perspectives. [**Result/Conclusion**] The extraction performance for the source attribute was the best, with Micro-F1 and Macro-F1 scores reaching at least 98.90% and 97.83%, respectively. This was followed by response options, usage instructions, and scoring rules, with item numbers and instructions showing moderate performance. Clinical explanations had the lowest performance, with Micro-F1 and Macro-F1 scores of 47.73% and 45.51%, respectively. ChatGPT-4 performed better overall, but the recall rate for some attributes was weaker than that of ChatGPT-3.5. Increases in measurement concept levels, dimensionality, number of items, and text length were found to reduce model performance. In summary, ChatGPT can efficiently assist in the structuring of medical scale resources, especially when dealing with simple scales.

Keywords: medical scale; text structuring; attribute extraction; large language model; zero-shot learning

Classification Number: TP391 G255

0 Introduction

Clinical scales are standardized measurement tools used in clinical research and practice, structured as questionnaires composed of carefully designed and selected items [1]. Tracking and obtaining the latest and best clinical scales has be-

come an important component of clinical activities [2]. However, a considerable portion of clinical scales exists in unstructured document formats, which creates computational burdens as clinical scale resources continue to grow. Therefore, there is an urgent need to adopt relevant methods to promote the structuring and intellectualization of clinical scale resources to facilitate computer processing.

An item refers to a single question, statement, or task (along with its standardized response options) used in clinical scales to assess specific concepts [3]. Items constitute the basic elements of clinical scales [4] and represent the smallest unit in clinical research and services. Establishing an item-centered environment for clinical scales that is discoverable, accessible, interoperable, and reusable has become the focus of research and practice in the intellectualization of clinical scale resources [5,6]. Efficiently extracting items and their attribute information from clinical scale documents to achieve structured representation of item information is a critical step in this process.

In recent years, methods based on Large Language Models (LLMs) [7] have become the state-of-the-art (SoTA) approach for information extraction tasks in natural language processing [8,9] and have been integrated into specialized application workflows such as entity annotation [10], knowledge question answering [11], knowledge graph construction [12], and knowledge generation [13]. Generative LLMs like ChatGPT possess powerful natural language understanding capabilities and can perform information extraction through dialogue using zero-shot and few-shot approaches [14-17], transforming the traditional pre-training plus fine-tuning paradigm. This reduces the pressure of data annotation and the cognitive burden of model usage, attracting attention in clinical information extraction and specialized application tasks [18-22]. These studies have further demonstrated from multiple perspectives—including economy, time, accuracy, and satisfaction—that LLMs can support clinical text resource processing and improve the efficiency and utilization potential of unstructured clinical data. However, to date, no evaluation studies have been reported on structured information extraction from clinical scale documents using LLMs.

This study focuses on processing Chinese clinical scale document structures, specifically targeting item information extraction tasks. We defined an item information extraction structure, designed prompt templates based on measurement concept levels, and implemented joint extraction of items and their attributes from clinical scale documents using the ChatGPT API. We further explored and evaluated the semantic processing capabilities of ChatGPT for multiple types of clinical texts. The main contributions are as follows: (1) We proposed a method for structured item information extraction from clinical scales using ChatGPT and zero-shot prompts, demonstrating the feasibility of deeply integrating general-purpose generative LLMs with clinical scale document structuring. (2) The designed structured item information extraction framework and zero-shot prompt templates can be reused for similar medical scale texts. (3) We conducted a fine-grained analysis of item structured infor-

mation extraction performance across different ChatGPT versions and various clinical scale contexts, providing references for strategy development in large-scale clinical scale text structuring. (4) We generated the first structured Chinese clinical scale item dataset, named CMedS-I, which contains 46 single-level measurement concept scales, 13 two-level measurement concept scales, 482 measurement concepts, 2,378 items, and eight types of item attributes, providing high-value corpora for related intelligent processing tasks.

1 Related Work

The goal of structured information extraction tasks is to identify attribute expressions in natural language text and present or store them according to specified structural templates. These attribute expressions can be entity names representing entities (e.g., "drug name" for drug entities), descriptive attributes representing features (e.g., "side effects" for drugs), or object/relationship attributes revealing relationships between entities (e.g., "indications" / "treatment" for drugs). For complex structural information such as knowledge representation models in knowledge graphs, pipeline methods [23] and joint learning methods [24,25] are commonly employed. The techniques used can be categorized into rule-based methods, traditional machine learning methods, and deep learning methods.

With breakthroughs in deep learning and the development of LLM fine-tuning technologies, LLM-based methods have become one of the mainstream approaches for structured information extraction in recent years. Current research on clinical text information extraction using LLMs primarily focuses on fundamental applications such as text annotation [8,26], text classification [27,28], and knowledge graph construction [22,29,30]. The extraction content mainly revolves around entities and their attributes and semantic relationships in domains such as diseases, symptoms, medications, and examinations. Evaluation methods include exact matching, fuzzy matching, and manual assessment, with metrics such as precision, recall, and F1-score. Research data primarily comes from public datasets like MIMIC, i2b2, SciERC, and PromptCBLUE [31], which contain electronic health records, medical literature, and question-answering texts. Studies commonly employ general-purpose generative LLMs such as ChatGPT, Claude, PaLM, and the LLaMA series, as well as their medically fine-tuned versions, focusing on prompt tuning to efficiently complete clinical information extraction tasks. For example, Agrawal et al. [22] used GPT-3 (zero-shot and one-shot) to extract medication information —including dosage, route, duration, and frequency—from the MIMIC dataset. Zhu et al. [30] evaluated entity relation extraction and reasoning capabilities using ChatGPT and GPT-4 with zero-shot and one-shot prompt strategies on SciERC, Re-TACRED, and DuIE2.0 datasets. These studies have not yet involved clinical scale text datasets or related information extraction tasks, but their research approaches and findings provide valuable references for this work.

In LLM applications, prompt engineering has gained popularity across various professional domains due to its low computational resource requirements and natural language interaction advantages, demonstrating excellent performance [32,33]. Zero-shot prompting and few-shot prompting—two fundamental approaches in prompt engineering—have become primary methods for solving real-world problems using LLMs [34]. Zero-shot prompting includes only the task description in the prompt template without any output examples, whereas few-shot prompting provides a small number of high-quality examples in the template. Typically, few-shot prompting yields better model performance but is limited by context length constraints. Zero-shot prompting, through instruction tuning, enables models to exhibit better generalization capabilities without requiring additional task-specific training data as examples in the template.

Furthermore, prompt design is crucial in prompt engineering. Based on task completion steps, relevant research employs prompt templates that can be broadly categorized into single-step prompts and multi-step prompts. Single-step prompts use one prompt to complete all task descriptions through a single dialogue round. For instance, Dunn et al. [35] designed prompts for joint named entity recognition and relation extraction to fine-tune GPT-3 and Llama-2, achieving single-round dialogue-based extraction of materials chemistry domain entities and their relationships from scientific texts. Multi-step prompts decompose tasks into multiple subtasks according to certain logic, designing separate prompts for each subtask and completing extraction through multi-round dialogue with the model [36,37]. Different prompt strategies have their respective advantages and disadvantages. The advantage of single-step prompts lies in their simple workflow and low economic cost, particularly when processing long texts. The disadvantage is that task description can be challenging, especially for complex tasks, and results are highly dependent on the model' s comprehension ability. Multi-step prompts can decompose complex tasks into simpler ones, reducing the model' s understanding burden, but suffer from complex workflows, heavy interaction overhead, time consumption, and high computational resource consumption or expenses due to repeated input-output operations. Overall, the choice of prompt engineering methods requires consideration of multiple factors, including task complexity, data characteristics, and time and resource costs.

2 Methodology

2.1 Overall Framework

This study explores the feasibility of extracting structured item information from clinical scale documents using ChatGPT and prompt engineering, focusing on commonly used clinical psychometric assessment scales. [Figure 1: see original paper] illustrates the research framework, which comprises four main components: (1) designing a clinical scale item information extraction framework to define target item attributes for prompt design reference; (2) selecting and preprocessing data to meet ChatGPT input requirements and form an ex-

perimental dataset; (3) designing prompt templates based on requirements to ensure task adaptability; and (4) calling different ChatGPT version APIs for experiments and evaluating the extraction performance of each item attribute while analyzing inter-version differences and potential influencing factors.

2.2 Item Structured Information Extraction Framework Design

Currently, no unified structured description framework exists for item information in clinical scales. Scale KRF provides a fine-grained definition of knowledge elements and semantic relationships in medical scale documents [38], offering references for scale resource knowledge representation and processing control. However, this framework describes scales rather than items and thus cannot be directly applied to item-centered structured information extraction. Based on Scale KRF, this study designed the Item Core Description Framework (ICDF). As shown in , ICDF converts content knowledge elements and measurement instruction knowledge elements related to items in Scale KRF into item attributes, considering the structural differences among various clinical scales. Eight attribute categories were defined: source, number, item instruction, response options, measurement concept, usage instructions, scoring rules, and clinical interpretation. Item instructions represent item entities; source and measurement concept are object attributes that can establish semantic relationships between items and scales or clinical concepts; number, response options, usage instructions, scoring rules, and clinical interpretation describe item measurement and usage characteristics.

Based on measurement objects and purposes, clinical scales have either single-level or two-level measurement concepts. In two-level measurement concept scales, the broader concept is called a domain, while the narrower concept is called a dimension or sub-domain; measurement of the former is achieved through the latter. In this study, the attribute name "measurement concept" is used directly in single-level measurement concept scale contexts. In two-level measurement concept scale contexts, to distinguish measurement concept levels, the attribute names "domain concept" and "dimension concept" are used. The attribute value examples in are derived from Item 1 in the single-level measurement concept scale "Self-Rating Depression Scale" [39].

2.3 Dataset Construction

As no relevant public datasets currently exist, this study selected commonly used clinical psychometric assessment scale documents to construct a research dataset. Following Huang et al.' s strategy for constructing standard datasets in clinical structured information extraction research [40], we generated a reference answer set as the basis for method evaluation.

2.3.1 Raw Data Selection We selected manuals of commonly used clinical psychological assessment scales from reference books as raw data, using the following criteria: (1) complete scale questionnaire appendices to ensure original

integrity of item information; (2) rich item attribute information in scale introduction texts; and (3) no image, audio, or video response items in scale content. We also considered structural differences in measurement concepts, including measurement concept levels, concept quantities, and item numbers.

2.3.2 Data Preprocessing: Model Input Data Generation First, we shortened the text length of each clinical scale document by manually removing paragraphs irrelevant to item structured extraction tasks to accommodate model input length requirements while retaining paragraph structure titles and original paragraph order. Second, we removed special characters and formatting characters (e.g., headers) from the documents. Finally, we checked text length. For documents still exceeding the maximum token limit, we prioritized removing irrelevant sentences from scale introduction texts and then considered removing scale content.

2.3.3 Reference Answer Set Generation Two researchers independently reviewed and corrected model output files based on the designed item knowledge extraction framework and preprocessed clinical scale documents to form structured item information reference answer sets for each scale document. For attributes with multiple possible mentions (e.g., "source" with scale names appearing in Chinese, English, abbreviations, or mixed forms), all mention forms in the original text were listed and separated by "@@@" . For item attributes without clear character boundaries, such as clinical interpretation, only reference answers were generated with the requirement that no important semantic fragments be omitted. The researchers conducted mutual checks—answers accepted by the other researcher were judged as consistent. For reference answers with disagreements, the original documents were consulted and discussed. Cohen' s Kappa Coefficient was used to check the consistency of review and correction results, yielding a coefficient of 72.66%, indicating high consistency and credible results [41]. Further analysis revealed that this was primarily affected by the "scoring rules" attribute, which had a Cohen' s Kappa coefficient of 55.24%, while other attributes had minimum Cohen' s Kappa values of 82.42%. The low consistency for "scoring rules" stemmed from inconsistent criteria between researchers. Scale documents often include two parts for scoring rules: the scoring method name and detailed scoring method description (e.g., "Responses use a Likert 5-point scoring method: 0=no occurrence or no impact, 1=mild impact, 2=moderate impact, 3=severe impact, 4=extremely severe impact"). One researcher believed that including either part was sufficient, while the other believed that if both appeared in the original text, the reference answer set should include both. After discussion, the latter opinion was adopted, and reference answers were revised accordingly, though model output files could contain either part.

2.4 Prompt Design

Prompts guide models to learn specific tasks, enabling rapid fine-tuning on small datasets for efficient task completion [10]. This study' s prompt design primarily

referenced the guiding principles and implementation strategies summarized by Andrew Ng et al. [42]. Following the iterative process shown in [Figure 2: see original paper]—which involves requirement specification, template design, preliminary experiments, error analysis, and revision of requirements and prompt templates based on results—we selected zero-shot prompting. The prompt template comprises four components: (1) role definition—specifying the model as a medical natural language processing model; (2) task description—extracting item attribute information from input text according to the given item attribute framework; (3) task requirements—extracting original text, outputting in JSON format, and specifying default values for unextracted attributes; and (4) input —preprocessed clinical scale documents.

Considering input document length, economic cost, and time consumption, this study adopted a single-step prompt strategy that simultaneously extracts the "item instruction" attribute representing the item entity and other attributes. Preliminary experiments revealed that using simple attribute list extraction structures made it difficult to clearly reveal semantic relationships between "item instruction" and other attributes, and extraction results contained numerous "NULL" values with severe "laziness" or "unwillingness to repeat output" phenomena. For example, in the same scale, some items extracted scale names for "source" while others returned empty results. To clearly reveal direct measurement relationships (dimension concepts) and indirect measurement relationships (domain concepts) between items and concepts in two-level measurement concept scales while addressing the model' s "laziness," the item attribute structure in the task description prompt element (see) was designed as follows: (1) Nested extraction structure: attributes with identical values across items were set as parent attributes that did not require repeated output, while attributes with different values were set as a group of child attributes. (2) Category-specific design: different extraction structures were used for different measurement concept levels. In the single-level measurement concept scale item information extraction structure (referred to as "single-level concept item information structure"), the measurement object attribute label is "measurement concept." In the two-level measurement concept scale item information extraction structure (referred to as "two-level concept item structure"), the measurement object attribute labels are subdivided into "dimension concept" and "domain concept," appearing at the same level to reduce task complexity and avoid performance degradation due to excessive nesting depth.

2.5 Experimental Configuration and Process

Experiments were conducted using the GPT-3.5 Turbo (16K-0613) and GPT-4 Turbo (gpt-4-0125-preview) API interfaces. We selected temperature sampling with temperature set to 0 to ensure more stable and reproducible model outputs, facilitating feedback-based prompt optimization. To maximize the model' s ability to reproduce original text content for information extraction tasks, the presence penalty was set to -2.0. All other parameters were set to default values.

Preprocessed text was used directly as input, with only one scale document input per API call.

2.6 Evaluation Method

Exact string matching, fuzzy matching, and manual evaluation are commonly used methods for assessing LLM performance in information extraction tasks. This study employed different accuracy determination schemes for different item attribute categories: (1) For attributes with strict boundaries, exact string matching was used, including source, number, item instruction, and measurement concept. If an attribute (e.g., source) had multiple mention forms in the text, model extraction results were judged accurate if they completely contained any mention form. (2) For scale item attributes without strict boundaries, manual "satisfaction" assessment was conducted through reading, marking "satisfactory" as "accurate" and unsatisfactory as "incorrect." The principle was that extraction results with no important semantic fragments omitted or that did not affect human understanding and usage compared to the original text were judged "satisfactory." This category included response options, usage instructions, scoring rules, and clinical interpretation.

For each scale document, item attribute information extraction performance was evaluated using precision (P), recall (R), and F1-score (F1). For item attribute a in the i-th scale document, let A be the set of instances extracted by the model for attribute a, and B be the set of all actual instances of attribute a contained in the i-th scale document. The precision, recall, and F1-score are calculated as follows:

$$P_i = \frac{TP}{TP + FP}, \quad R_i = \frac{TP}{TP + FN}, \quad F1_i = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}$$

where TP is the number of accurately extracted attribute instances in A , FP is the number of incorrectly extracted attribute instances in A , and FN is the number of attribute instances in B that were not extracted.

For overall extraction performance evaluation of an item attribute a across all scale documents, we referenced question-answering information extraction evaluation methods, treating each scale as a class and evaluating from both micro and macro levels:

1) Micro-precision (Micro-P), micro-recall (Micro-R), and micro-F1 (Micro-F1) are calculated as:

$$\text{Micro-P} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)}, \quad \text{Micro-R} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)}, \quad \text{Micro-F1} = 2 \cdot \frac{\text{Micro-P} \cdot \text{Micro-R}}{\text{Micro-P} + \text{Micro-R}}$$

2) Macro-precision (Macro-P), macro-recall (Macro-R), and macro-F1 (Macro-F1) are calculated as:

$$\text{Macro-P} = \frac{1}{n}\sum_{i=1}^{n} P_i, \quad \text{Macro-R} = \frac{1}{n}\sum_{i=1}^{n} R_i, \quad \text{Macro-F1} = 2 \cdot \frac{\text{Macro-P} \cdot \text{Macro-R}}{\text{Macro-P} + \text{Macro-R}}$$

where n is the number of scale documents.

3 Results and Analysis

3.1 Dataset Characteristics

The CMedS-I dataset contains 46 single-level measurement concept scales and 13 two-level measurement concept scales, all multi-dimensional, totaling 164,032 Chinese characters and 240,302 tokens. The dataset includes 482 measurement concepts and 2,378 items. Specifically, single-level measurement concept scales contain 261 measurement concepts and 1,529 items, while two-level measurement concept scales contain 43 domain concepts, 178 dimension concepts, and 849 items. Since response options are identical within a single scale, the number of response option sets equals the number of scales. All items have numbers. See for details. Not all scale documents contain item clinical interpretations. In this dataset, 35 scale documents have explicit item clinical interpretations, while 24 do not. As shown in [Figure 3: see original paper], the dimension levels (i.e., number of measurement concepts) of selected scales range from 2 to 71, mostly concentrated between 2-10; the number of items per scale ranges from 10 to 148; and text lengths are primarily between 1.5k and 3.5k characters.

3.2 Experimental Results and Analysis

and present evaluation results for GPT-3.5 Turbo and GPT-4 Turbo extraction performance across different measurement concept structure scales from micro and macro perspectives, respectively.

3.2.1 Overall Extraction Effect Analysis As shown in and , both ChatGPT-3.5 and ChatGPT-4 exhibited significant differences in precision, recall, and F1 values across different item attributes, particularly in two-level measurement concept scale contexts. [Figure 4: see original paper] demonstrates that overall, source performance was the best, followed by response options, usage instructions, and scoring rules, with item numbers and instructions showing moderate performance, and measurement concepts, dimension concepts, domain concepts, and clinical interpretations performing the worst.

In single-level measurement concept scale contexts, source attribute extraction achieved the best performance, with Micro-F1 and Macro-F1 scores reaching at least 98.90% and 97.83%, respectively. Item instruction performance was moderate, with Micro-F1 and Macro-F1 scores of 66.75% and 77.12%, respectively. Although clinical interpretation performed the worst, its Micro-F1 and Macro-F1 scores still reached 47.73% and 45.51%, respectively.

In two-level measurement concept scale contexts, the overall performance ranking of attributes remained consistent, but specific performance differed from single-level contexts. Except for source and response options, all attributes showed varying degrees of decline in precision, recall, and F1 values under ChatGPT-3.5. Clinical interpretation, domain concept, and dimension concept extraction performance declined significantly. Micro-F1 and Macro-F1 for clinical interpretation decreased from 59.22% and 54.48% to 8.11% and 16.70%, respectively. Dimension concept and domain concept Micro-F1 scores were below 18% and Macro-F1 scores below 24% under ChatGPT-3.5, far lower than measurement concept extraction performance in single-level measurement concept contexts.

These differences may be related to attribute value constraints and output structure design in prompt templates. In the prompt template, for attributes such as source, response options, usage instructions, and scoring rules—which generally have identical values across different items in the same scale—only one output was required. For measurement concepts, dimension concepts, domain concepts, and clinical interpretations—which have different values across items in multi-dimensional clinical scales selected for this study—repeated output was required in the prompt template to reflect the correspondence between these attributes and items (i.e., "item instruction" values), reducing subsequent item-centered structuring burden. Further error analysis revealed that low precision was related to "hallucination" phenomena [43], where ChatGPT generated new attribute information beyond the input text, particularly in ChatGPT-4. This confirms previous findings of "insufficient constrained output ability" [12]. Additionally, ChatGPT-4' s low recall may be related to its deduplication mechanism.

3.2.2 Different Model Version Performance Analysis Overall, as shown in and , ChatGPT-4 demonstrated outstanding precision across item attributes but severely low recall compared to its precision, showing obvious performance imbalance. ChatGPT-3.5 achieved precision comparable to ChatGPT-4 for some attributes, particularly in single-level measurement concept contexts, with more balanced performance. In two-level measurement concept scale contexts, ChatGPT-4 outperformed ChatGPT-3.5 in both precision and recall. Specifically:

In terms of precision: Except for scoring rules in two-level measurement concept contexts, ChatGPT-4 outperformed ChatGPT-3.5. In complex two-level measurement concept scale contexts, ChatGPT-4 achieved 100% precision for source, number, and item instruction attributes, with the lowest Macro-P for clinical interpretation still reaching 64.08%.

In terms of recall: ChatGPT-3.5 outperformed ChatGPT-4 for number and item instruction extraction in both Micro-R and Macro-R. Measurement concept extraction was only slightly better than ChatGPT-4 in single-level measurement concept scale documents but significantly worse in two-level measurement concept scale documents. ChatGPT-3.5 performed worse than ChatGPT-4 for

response options, usage instructions, scoring rules, and clinical interpretation.

In terms of Micro-F1 and Macro-F1: As shown in [Figure 4: see original paper], ChatGPT-3.5' s source attribute extraction performance was comparable to ChatGPT-4; item instruction extraction outperformed ChatGPT-4; number and clinical interpretation extraction were only slightly better than ChatGPT-4 in single-level measurement concept contexts, with clinical interpretation extraction significantly weaker in two-level measurement concept contexts; response options, usage instructions, and scoring rules extraction were all weaker than ChatGPT-4.

This again demonstrates that different ChatGPT versions perform differently on the same clinical scale item structured information extraction task, and ChatGPT-4 is not always superior to ChatGPT-3.5, consistent with findings in other clinical text processing contexts [44].

3.2.3 Analysis of Model Performance Under Different Scale Characteristics As shown in and , measurement concept levels in scales affected both precision and recall for both model versions, particularly for item instruction, measurement concept (dimension concept and domain concept in two-level scales), and clinical interpretation attributes. This section further analyzes extraction performance for these three attributes under different numbers of items, dimension levels, and text lengths based on Micro-P and Micro-R metrics, using the better-performing single-level measurement concept scale context to identify other scale characteristics that may affect model performance.

(1) Analysis of Model Performance Under Different Numbers of Items

[Figure 3: see original paper] shows that 93.48% (43/46) of single-level measurement concept scales have fewer than 80 items. As shown in [Figure 5: see original paper], within this range, as the number of items in a scale increases: (1) For precision, item instruction precision remains relatively stable, while other attributes show obvious fluctuations; measurement concept fluctuations are relatively larger under ChatGPT-3.5, and clinical interpretation fluctuations are relatively larger under ChatGPT-4. (2) For recall, item instruction recall remains relatively stable under ChatGPT-3.5 but shows a downward trend under ChatGPT-4; recall for other attributes shows a downward trend amid fluctuations.

(2) Analysis of Model Performance Under Different Dimension Levels

[Figure 3: see original paper] shows that single-level measurement concept scale dimension numbers are primarily concentrated below 7 (approximately 82.61%), with one scale each at 7, 8, and 11 dimensions, two scales at 10 dimensions, and three scales at 20 dimensions. As shown in [Figure 6: see original paper], for scales with fewer than 7 dimensions, as dimension number increases: (1) For precision, item instruction precision remains relatively stable, measurement concept shows obvious fluctuations, and clinical interpretation remains relatively

stable except for relatively large fluctuations at 5 dimensions under ChatGPT-4. (2) For recall, item instruction recall remains relatively stable under ChatGPT-3.5 but shows a downward trend under ChatGPT-4; recall for other attributes shows a downward trend amid fluctuations.

(3) Analysis of Model Performance Under Different Context Lengths

[Figure 3: see original paper] shows that 89.13% (41/46) of single-level measurement concept scales have input text lengths between 1.5k and 3.5k characters. As shown in [Figure 7: see original paper], within this range, as text length increases: (1) For precision, item instruction precision remains relatively stable, while other attributes show obvious fluctuations; measurement concept shows an upward trend under ChatGPT-3.5 but a downward trend under ChatGPT-4. (2) For recall, item instruction recall remains relatively stable under ChatGPT-3.5 but shows a downward trend under ChatGPT-4; recall for other attributes shows an overall downward trend.

3.3 Analysis of Prompt Effects Using Different Attribute Explanation Patterns

Given that LLM text processing performance may be affected by their understanding of professional terminology—i.e., the level of detail in attribute explanation prompts may impact output results—this study randomly selected five single-level measurement concept scales and conducted experiments using ChatGPT-3.5 to compare output effects under three modes: no explanation, simple explanation (providing only attribute definitions), and detailed explanation (including both attribute definitions and value specifications). [Figure 8: see original paper] presents a heatmap of extraction performance under different explanation modes. Results show that detailed explanation prompts outperformed no explanation and simple explanation overall in Macro-P, Macro-R, and Macro-F1. This finding indicates that providing detailed attribute explanations can significantly improve LLM text processing capabilities. However, we also observed that for certain attributes (e.g., usage instructions and clinical interpretation), simple explanation performed worse than no explanation, possibly because additional explanation information caused interference in model understanding and processing. Meanwhile, adding value specifications provides more specific contextual information, which positively impacts extraction performance and helps more accurately understand and process professional terminology. Therefore, when designing attribute explanation prompts for LLMs, the level of detail for each attribute explanation should be considered to achieve optimal text extraction performance and improve overall model performance.

4 Conclusions and Perspectives

Addressing the current lack of evaluation studies on ChatGPT-based structured information extraction in clinical scale text contexts and the practical needs for intelligent development of clinical scale resources, this study focused on item

knowledge elements in clinical scales and proposed a method for extracting structured item information from clinical scale documents using ChatGPT and zero-shot prompts, achieving joint extraction of items and multiple attribute information. The method first defines a preset item structured information extraction framework by defining core item attributes, then designs prompt templates accordingly. The prompt templates consider multiple factors, including structural differences in clinical scale content, revelation of semantic relationships among item attributes, and ChatGPT' s inherent "laziness" phenomenon. Using clinical psychological and behavioral scales as examples, we constructed a self-built multi-type clinical scale dataset and evaluated the proposed method using ChatGPT-3.5 and ChatGPT-4 open APIs. The main conclusions are:

(1) ChatGPT can assist in structured item information extraction from Chinese unstructured clinical scale texts, especially in single-level measurement concept scale contexts. When processing single-level measurement concept scale documents, although clinical interpretation performed the worst overall, its Micro-F1 and Macro-F1 scores still reached 47.73% and 45.51%, respectively; the second-worst measurement concept achieved Micro-F1 and Macro-F1 scores of 57.81% and 62.73%, respectively.

(2) Extraction performance varies significantly across different item attributes. In this study, single-output attributes such as source, response options, usage instructions, and scoring rules performed significantly better than repeated-output attributes such as number, item instruction, measurement concept, dimension concept, domain concept, and clinical interpretation.

(3) Different ChatGPT versions perform differently on the same clinical scale item structured information extraction task, and ChatGPT-4 is not always superior to ChatGPT-3.5 as expected. In terms of F1 scores, ChatGPT-3.5' s source attribute extraction performance was comparable to ChatGPT-4, and its item instruction attribute extraction outperformed ChatGPT-4.

(4) Scale structural and text characteristics affect ChatGPT item structured information extraction performance to varying degrees. Both ChatGPT-3.5 and ChatGPT-4 performed better overall in single-level measurement concept clinical scale contexts than in two-level measurement concept scale contexts. Increases in measurement concept numbers, item numbers, and text length reduced model performance.

The limitations of this study include that, due to dataset construction and manual evaluation workload, the types of clinical scales and item attributes included were not sufficiently rich (e.g., scales with image-based response items were not covered). Future work should incorporate more types of clinical medical scales and evaluate them in newer ChatGPT versions and more LLMs. Additionally, inspired by these results, future efforts will continue to explore strategies for im-

proving model extraction capabilities through optimizing extraction structures in prompt templates, adding attribute value constraint specifications, and error analysis-oriented self-verification.

Overall, this study further explores ChatGPT' s semantic processing capabilities for multiple types of clinical texts and demonstrates the feasibility of deeply integrating general-purpose generative LLMs with clinical scale document structuring and organization. The results can provide references for model usage strategy development in clinical scale structuring projects. The generated CMedS-I dataset—the first structured Chinese clinical scale item dataset—can provide high-value corpora for related research and downstream tasks such as scale knowledge graph construction, knowledge computation, and knowledge retrieval.

References

[1] SHI R, GUO A M. Research methods of general practitioners[M]. Beijing: People' s Medical Publishing House, 2017: 211-219.

[2] ZHU M, HONG R H, YANG T, et al. The Efficacy of measurement based care for depressive disorders: Systematic review and meta-analysis of randomized controlled trials[J]. The journal of clinical psychiatry, 2021, 82(5): 21r14034.

[3] U.S. FOOD AND DRUG ADMINISTRATION. Patient-reported outcome measures: use in medical product development to support labeling claims[EB/OL]. [2023-9-15]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims.

[4] DEVELLIS R F. Scale development: Theory and applications[M]. 4th ed. Los Angeles: SAGE, 2017.

[5] CELLA D, YOUNT S., ROTHROCK N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years[J]. Medical care, 2007, 45(5 Suppl 1): S3-S11.

[6] NLM. NIH CDE repository[EB/OL]. [2024-01-4]. https://cde.nlm.nih.gov/form/search.

[7] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[J]. arXiv:2303.18223, 2023.

[8] GOEL A, GUETA A, GILON O, et al. LLMs accelerate annotation for medical information extraction[C]//Machine learning for health (ML4H). PMLR, 2023: 82-100.

[9] WAN Z, CHENG F, MAO Z, et al. GPT-RE: In-context learning for relation extraction using large language models[J]. arXiv preprint arXiv:2305.02105, 2023.

[10] YANG D G, HUANG J T. Named entity recognition method of large language model for medical question answering system[J/OL]. Computer engineering: 1-7[2024-03-30]. https://doi.org/10.19678/j.issn.1000-3428.0068400.

[11] Karan S, Tao T, Juraj G, et al. Towards expert-level medical question answering with large language models[J]. arXiv preprint arXiv:2305.09617, 2023.

[12] FREY J, MEYER L P, ARNDT N, et al. Benchmarking the abilities of large language models for RDF knowledge graph creation and comprehension: How well do LLMs speak turtle?[J]. arXiv preprint arXiv:2309.17122, 2023.

[13] KUMAR A H S. Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain[J]. Biology, engineering, medicine and science reports, 2023, 9(1): 1-5.

[14] MANNING C D. Human language understanding & reasoning[J]. Daedalus, 2022, 151(2): 127-138.

[15] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[16] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1-35.

[17] HAN X, ZHANG Z, DING N, et al. Pre-trained models: past, present and future[J]. arXiv preprint arXiv:2106.07139, 2021.

[18] CHENG P, Xi Y, AOKUN C, et al. A study of generative large language model for medical research and healthcare[J]. arXiv preprint arXiv:2305.13523, 2023.

[19] MA W R, GONG M C, DAI H, et al. A comprehensive review of the applications of large language models in clinical medicine with ChatGPT as a representative[J]. Journal of medical informatics, 2023, 44(7): 9-17.

[20] Monica A, Stefan H, Hunter L, et al. Large language models are few-shot clinical information extractors[C]//The 2022 conference on empirical methods in natural language processing, 2022: 1998-2022.

[21] PENG C, YANG X, CHEN A, et al. A study of generative large language model for medical research and healthcare[J]. NPJ digital medicine, 2023, 6(1): 210.

[22] AGRAWAL M, HEGSELMANN S, LANG H, et al. Large language models are few-shot clinical information extractors[J]. arXiv preprint arXiv:2205.12689, 2022.

[23] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[J]. arXiv preprint arXiv:1601.00770, 2016.

[24] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert systems

with applications, 2018, 114: 34-45.

[25] LUAN Y, HE L, OSTENDORF M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[J]. arXiv preprint arXiv:1808.09602, 2018.

[26] YANG B, SUN X H, DANG J Y, et al. Named entity recognition method of large language model for medical question answering system[J]. Journal of frontiers of computer science and technology, 2023, 17(10): 2389.

[27] SUN X, LI X, LI J, ET A L. Text classification via large language models[J]. arXiv preprint arXiv:2305.08377, 2023.

[28] ONG J, KEDIA N, HARIHAR S, et al. Applying large language model artificial intelligence for retina international classification of diseases (ICD) coding[J]. Journal of medical artificial intelligence, 2023, 6: 21.

[29] WU L I, LI G. Zero-shot construction of Chinese medical knowledge graph with ChatGPT[C]//2023 IEEE international conference on medical artificial intelligence (MedAI). IEEE, 2023: 278-283.

[30] ZHU Y, WANG X, CHEN J, et al. LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities[J]. arXiv preprint arXiv:2305.13168, 2023.

[31] ZHU W, WANG X, CHEN M, et al. Overview of the PromptCBLUE shared task in CHIP2023[J]. arXiv preprint arXiv:2312.17522, 2023.

[32] LI B, FANG G, YANG Y, et al. Evaluating ChatGPT' s information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness[J]. arXiv preprint arXiv:2304.11633, 2023.

[33] WEI X, CUI X, CHENG N, et al. Zero-shot information extraction via chatting with ChatGPT[J]. arXiv preprint arXiv:2302.10205, 2023.

[34] MIALON, G. DESSÌ R, LOMELI M, et al. Augmented language models: A survey[J]. arXiv preprint arXiv:2302.07842, 2023.

[35] DUNN A, DAGDELEN J, WALKER N, et al. Structured information extraction from complex scientific text with fine-tuned large language models[J]. arXiv preprint arXiv:2212.05238, 2022.

[36] ZHENG C, GUO C Y. A pattern-first pipeline approach for entity and relation extraction[J]. Neurocomputing, 2022(494): 182-191.

[37] XUE L L, ZHANG D, DONG Y X, et al. AutoRE: Document-level relation extraction with large language models[J]. arXiv preprint arXiv:2403.14888, 2024.

[38] SUN H X, HAO J, GUO Z, et al. Construction of a fine-grained knowledge element-based framework for knowledge representation in medical scale documents[J]. Digital library forum, 2023, 19(12): 86-98.

[39] DAI X Y. Handbook of commonly used psychological assessment scales: Revised edition[M]. Beijing: People' s Military Medical Publishing House, 2015: 150-154.

[40] HUANG J W, YANG D H M, RONG R C, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes[J]. NPJ digital medicine. 2024(7): 106.

[41] LANDIS J R, KOCH G G. The measurement of observer agreement for categorical data[J]. Biometrics, 33(1): 159-174.

[42] ANDREW N G, LSA engineering developers[EB/OL]. [2023-12-25]. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/.

[43] YUAN W, NEUBIG G, LIU P. Bartscore: Evaluating generated text as text generation[C]//The 34th international conference on neural information processing systems, 2021, 34: 27263-27277.

[44] CHEN J, LIU L, RUAN S, et al. Are different versions of ChatGPT' s ability comparable to the clinical diagnosis presented in case reports? A descriptive study[J]. Journal of multidisciplinary healthcare, 2023, 16: 3825-3831.

Author Contributions

Jie Hao: Designed the research protocol, constructed the dataset, analyzed experimental results, drafted the manuscript, and revised the final version.
Zhiqiang Mo: Constructed the dataset, conducted experiments, and analyzed experimental results.
Haixia Sun: Designed the research protocol, constructed the dataset, analyzed experimental results, drafted the manuscript, and revised the final version.
Zhenli Chen: Constructed the dataset and analyzed experimental results.
Jiao Li: Revised the research protocol and modified the manuscript.

Note: Figure translations are in progress. See original paper for figures.