Original Paper

An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study

Lei Wang[1,2*], MS; Yinyao Ma[3*], MS, MD; Wenshuai Bi[1], MS; Hanlin Lv[1], MD, PhD; Yuxiang Li[1,2], PhD

[1]BGI Research, Wuhan, China

[2]Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research, Shenzhen, China

[3]Department of Obstetrics, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, China

[*]these authors contributed equally

Corresponding Author:
Hanlin Lv, MD, PhD
BGI Research
1-2F, Building 2, Wuhan Optics Valley International Biomedical Enterprise Accelerator Phase 3.1
No 388 Gaoxin Road 2, Donghu New Technology Development Zone
Wuhan, 430074
China
Phone: 86 18707190886
Email: lvhanlin@genomics.cn

Abstract

Background: The study of disease progression relies on clinical data, including text data, and extracting valuable features from text data has been a research hot spot. With the rise of large language models (LLMs), semantic-based extraction pipelines are gaining acceptance in clinical research. However, the security and feature hallucination issues of LLMs require further attention.

Objective: This study aimed to introduce a novel modular LLM pipeline, which could semantically extract features from textual patient admission records.

Methods: The pipeline was designed to process a systematic succession of concept extraction, aggregation, question generation, corpus extraction, and question-and-answer scale extraction, which was tested via 2 low-parameter LLMs: Qwen-14B-Chat (QWEN) and Baichuan2-13B-Chat (BAICHUAN). A data set of 25,709 pregnancy cases from the People's Hospital of Guangxi Zhuang Autonomous Region, China, was used for evaluation with the help of a local expert's annotation. The pipeline was evaluated with the metrics of accuracy and precision, null ratio, and time consumption. Additionally, we evaluated its performance via a quantified version of Qwen-14B-Chat on a consumer-grade GPU.

Results: The pipeline demonstrates a high level of precision in feature extraction, as evidenced by the accuracy and precision results of Qwen-14B-Chat (95.52% and 92.93%, respectively) and Baichuan2-13B-Chat (95.86% and 90.08%, respectively). Furthermore, the pipeline exhibited low null ratios and variable time consumption. The INT4-quantified version of QWEN delivered an enhanced performance with 97.28% accuracy and a 0% null ratio.

Conclusions: The pipeline exhibited consistent performance across different LLMs and efficiently extracted clinical features from textual data. It also showed reliable performance on consumer-grade hardware. This approach offers a viable and effective solution for mining clinical research data from textual records.

(J Med Internet Res 2024;26:e54580) doi: 10.2196/54580

KEYWORDS

clinical data extraction; large language models; feature hallucination; modular approach; unstructured data processing

Introduction

Clinical text data have been widely recognized in data research due to their inclusion of multisource information [1,2] (eg, patient subjective statements, past objective facts, doctors' diagnostic processes, and summary records). Extracting useful information from text data could serve as a crucial supplement to the study of disease progression; it could complement objective indicators dependent on laboratory tests and examinations [3], which has consistently been a hot research topic [4,5].

Historically, methods for text data extraction mainly include the following:

- Manual annotation: scales are designed based on clinical and research experience, followed by manual field extraction [6-8].
- Rule extraction: concepts from established knowledge base, such as *International Classification of Diseases, Tenth Revision* [9], are used for concept term extraction. This process is typically based on similarity algorithms and manual assistance to extract terms and their attributes (eg, negations and dependency relationships) [10].
- Named entity recognition or natural language processing algorithms: supervised learning methods, such as pretrained models like T5 [11], Bidirectional Encoder Representations from Transformers (BERT) [12], and BERT's variants [13-15], with manual annotation to enhance semantic comprehension capabilities [16,17].

The task of extracting features from vast unstructured text presents itself as a daunting, labor-intensive, and time-consuming endeavor [18], for the following reasons:

- It is challenging to determine the dimension of extracted features initially, and from another perspective, confining the feature dimension means constraining the research scope from the outset [19].
- Given the inherent subjectivity and potential biases of recording subjects, solely relying on algorithms without annotation typically results in low accuracy and recall [20].
- Achieving higher accuracy with a broader feature scope, and the required human effort involved, is typically nonlinear [4], and the difficulty becomes apparent when confronted with massive real-world data.

The advent of large language models (LLMs) has paved a new path for the dilemma in clinical text extraction [21-23]. In the realm of natural language understanding research, generative large models, represented notably by ChatGPT [24] since 2022, have achieved unimaginable capabilities in semantic dimensions, leveraging the emergent intelligence from vast parameter scales. However, there are numerous considerations and limitations in their application, as follows:

- High-performing LLMs, such as OpenAI ChatGPT and Google Bard [25,26], are currently not open source, and patient data need to be submitted to their platform for analysis, presenting security challenges [27,28].
- Open-source LLMs with high intelligence generally require a large number of parameters (10-100 billion), which are hard to support on consumer-level graphics processing units (GPUs) [29].
- Low-parameter (around 10 billion) LLMs, typically require multistrategy support when dealing with tasks in certain vertical segments [30] (eg, fine-tuning, knowledge base or knowledge graph support, complex Chain of Thought (CoT) [31] along with its derivatives, and even global training) and are accompanied by various anomaly issues, including feature hallucination.

Although the application of LLM faces various potential limitations and challenges, as mentioned above, the foundational entity extraction and understanding capabilities of LLMs can still be used for low-cost extraction of clinical text data through meticulous prompt design, guidance combining CoT, and standardized examples [30,32].

In this study, we aimed to extract valuable features from a series of given patient admission records, which include the chief complaint and the medical histories. In light of this task, we introduced a modular LLM approach, which divides the entire extraction path into several smaller steps, with each modular LLM handling these basic steps automatically. We adopted the core idea of LLM agents [30] and self-consistency with CoT [33].

To experiment with this approach, we implemented 2 low-parameter LLMs in a local environment and compared their performances within a retrospective cohort of pregnancy to provide a reference that future researchers might draw upon.

Methods

Study Preparation

Data Sources

In this study, the text corpus was compiled from two primary sources:

1. Chief complaints and medical histories, exemplified in Multimedia Appendix 1, were extracted from inpatient admission records of an established cohort at the People's Hospital of Guangxi Zhuang Autonomous Region in China. The established cohort for the preeclampsia risk study consisted of 25,709 pregnancies that received prenatal care between the 11th and 13th weeks of gestation from April 2012 to September 2021.
2. Clinical practice guidelines consisted of the 2018 guidelines from the American College of Obstetricians and Gynecologists [34] and the 2019 guidelines from the National Institute for Health and Care Excellence [35].

To ensure linguistic consistency, the entire corpus was maintained in Chinese.

Model Deployment

We deployed 2 most exemplary LLMs in the Chinese domain until September 2023 in an intranet security environment independently: Qwen-14B-Chat (QWEN) [36] and Baichuan2-13B-Chat (BAICHUAN) [37]. In the environment, the server cluster used NVIDIA DGX-A100 (2×40 G) GPU nodes. The QWEN used 29 GB of storage and 27 GB of GPU memory, while the BAICHUAN used 26 GB of storage and 28.9 GB of GPU memory. Both models operated solely on physically isolated GPUs, and access was facilitated through the OpenAI [38] format and FastChat [39]. The LLMs were built upon PyTorch 2.0, with the temperature set to 0 and max_token adjusted task by task.

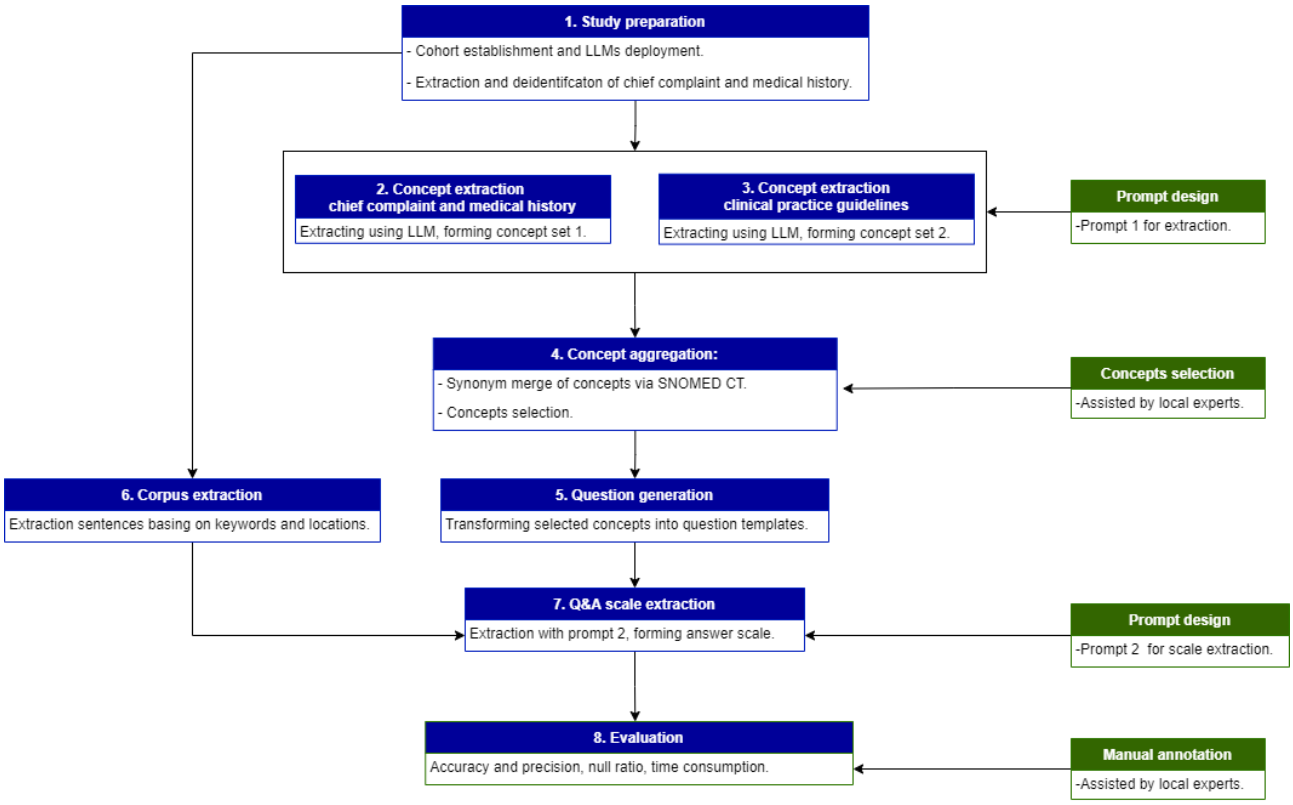
Experimental Path

In this study, we have introduced an approach that autonomously extracts valuable textual features. Diverging from traditional LLM applications, we used an "external-COT" strategy, dividing

the process into several controllable steps, as illustrated in Figure 1.

The extraction approach could be divided into four parts: (1) concept preparation, that is, extracting existing concepts from the corpus and selecting concerning concepts; (2) corpus preparation, that is, deidentifying raw data and preparing the corpus in accordance with the selected concept; (3) prompt design for different LLM tasks; (4) question-and-answer (Q&A) scales, that is, transforming concepts into question templates and extracting corresponding scales by LLMs.

Figure 1. The flowchart of extraction approach. LLM: large language model; Q&A: question and answer; SNOMED CT: Systematised Nomenclature of Medicine Clinical Term.



Prompt Design

The design of prompt templates is fundamental to efficient and accurate extraction. Prior to processing the entire data set, an initial evaluation was conducted on 100 observations to assess the effectiveness of the templates, allowing for continuous refinement of prompt strategies and orientations. An appropriate template was defined based on the following criteria: (1) absence of redundant content generation, (2) consistent and uniform efficiency, and (3) infrequent occurrence of feature hallucination.

We adopted a 4-paragraph structure, referring to the prompt engineering suggestions of QWEN and BAICHUAN, as follows:

1. Context section: defines the role and task, provides a basic understanding, and establishes a behavioral baseline for the model.
2. Instruction section: outlines the execution steps, uses the CoT methodology, and provides examples to ensure guided model operation.
3. Input data section: manages various inputs to meet diverse information needs.
4. Output indicator section: specifies the output format and standards, setting clear expectations for the output.

To avoid input bias, the prompt templates for QWEN and BAICHUAN were maintained without any modifications. In addition, we conducted experiments using 100 observations at different levels of concurrency to select the most optimal configuration.

Concept Extraction and Aggregation

We initially extracted all discernible concepts from chief complaints and medical histories using LLMs with a designed prompt 1, and concepts were retained only with a manifestation frequency exceeding 5% occurrences. To reduce potential attention bias and expand the range of identified concepts, we also included concepts from clinical practice guidelines related to preeclampsia, particularly the American College of Obstetricians and Gynecologists 2018 guidelines and the National Institute for Health and Care Excellence 2019 guidelines.

As we defined in prompt 1, the extracted concepts were formatted using the Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) vocabulary within the Clinical Findings and Observations domain [40].

To mitigate potential output errors from LLMs, such as concepts not belonging to the Clinical Findings and Observations domain, or even errors outside of the SNOMED CT vocabulary, we

implemented a rule-based matching approach to filter extraction inaccuracies.

Furthermore, in this research, we aimed to extract concepts with diverse semantic expressions (including diagnoses, various medical histories, symptoms, observations, interventions, and types of examination). To accomplish this, local experts manually filtered out concepts embedded in structured text, such as dates or numbers.

Question Generation

After the extraction and aggregation of concepts, they were transformed into specific questions by LLMs as question templates for subsequent data extraction. In this section, we leveraged ChatGPT4.0 as a question generator to produce a basic set of questions, which were then refined by local experts for specificity based on its performance across 100 observations.

Q&A Scale Extraction

To avoid contextual and temporal event confusion leading to incorrect responses (eg, confusing current medical history with a past medical history or confusing the patient's medical history with that of family members), we preextracted the corpus using two strategies: (1) based on the position of the question templates and (2) based on the sentence containing the concepts. The extracted corpus was then labeled with the corresponding question templates for the subsequent extraction of Q&A scales.

The refined corpus, combined with corresponding question templates, guided a systematic extraction process with 2 LLMs, forming Q&A scales for further application.

Each question probed the LLMs, and the extracted sentences formed the basis of the generated responses. This approach enabled a logical mapping of questions to relevant text, ultimately improving the accuracy and efficiency of feature extraction.

Evaluation

Given the practical constraints and the objective of minimizing manual intervention, it was unfeasible to validate all answer scales individually across a Q&A space containing 68 questions and 25,709 observations. Therefore, a 3-fold assessment strategy was developed, as explained in the sections that follow.

Accuracy and Precision

A subset of 1500 observations chosen at random was manually annotated in collaboration with local experts, serving as the gold standard. The precision of positive identifications by both LLMs was assessed against a specified benchmark.

Null Ratio

The null ratio of both LLMs was independently measured across all 25,709 observations. Empty or meaningless outputs (symbols and gibberish) were identified as null outputs, and the null ratio was then calculated as the proportion of such responses to the total.

Time Consumption

The efficiency of the extraction process was evaluated by measuring the time taken by the 2 LLMs to respond to the questions across all 25,709 observations.

Ethical Considerations

The study was approved by the People's Hospital of the Guangxi Zhuang Autonomous Region in China (KT-KJT-2021-67). The requirement for informed consent was waived, due to the retrospective nature of the study, and all clinical data were deidentified and anonymized.

Results

Path Decomposition Overview

Prompt Template

Through trials with the prompt template on 100 observations, we selected the template that demonstrated optimal consistency, as shown in Figure 2.

Figure 2. Prompt templates of information extraction and question-and-answer scales. A 4-paragraph structure was implemented for a prompt design using a few-shot Chain-of-Thought prompting. The original version, written in Chinese, was translated into English. (A) Extracting concepts and terms; (B) scaling questions. As both deployed large language models were pretrained with the *International Classification of Diseases, Tenth Revision* corpus, they could directly engage in concept extraction.

(A)

```
Prompt=f'''
You are a stringent information extractor, you should work according to
operations meeting requirements, and referring to examples to extract
clinical medical concepts and terms from the text. You are very strict; you
will absolutely not associate or produce information that is not in the
original text.
===
#### Text: {}
===
#### Operations:
1. Decompose the text sentence by sentence, breaking it down according to the
independent meaning of each sentence. The number of breakdowns should be as
many as possible, forming individual information blocks.
2. If the word count of an information block exceeds 10 words, it needs to be
further decomposed until the length of each information block does not exceed
10 words.
3. For each information block, summarize its core meaning into a keyword or
concept.
4. Check these core concepts or keywords, ensuring that the length of each
concept does not exceed 10 words, and ensure they are semantically accurate.
5. List the standardized core concepts or keywords, and use '\\n' to separate
them.
===
#### Requirements:
Concepts and terms belong to diseases, symptoms, drugs, treatment methods,
test items, examination items, medical history, personal information. During
the execution process, pay attention to identifying synonyms or various
expressions of medical terms. Do not output details of the original text.
===
#### Output format:
Only need to output the result of step five. Start with "```row{}"
===
#### Sample:
Example 1:
Text: "The patient was diagnosed with Type 2 diabetes in 2021 and is
currently controlling blood sugar by using metformin and performing aerobic
exercise."
Output: " row1: Type 2 diabetes\\n Metformin\\n Aerobic exercise".
Example 2: "The patient is gravida 3, para 1, 39 weeks + 1 day pregnant."
Output: " row2: Gravida\\n Para\\n Gestational week"
'''
```

(B)

```
Prompt=f'''
You are a stringent clinical data scientist, working according to
operational steps, meeting requirements, referencing examples, extracting
and understanding medical information related to legs from given documents
in order to answer queries about it. Information not present in the original
text will absolutely not be inferred or generated.
===
#### Text: {}
===
#### Question: {}
===
#### Operations:
1. Identify the key term in the question.
2. Check if the key term appears in the text.
3. Determine whether the key term is negated in the text.
4. Output according to the requirements
===
===
#### Requirements: {}
===
#### Output format:
Begin with "```row{}" and end with "```"
===
#### Sample:
Example 1:
Question: "How is the patient's menstrual flow?"
Requirements: "Choose only: Light/Moderate/Heavy/Unknown"
Document: "...painless periods, moderate flow..."
Output: " row1: Moderate"
Example 2:
Question: "5. Does the patient have a history of alcohol consumption?"
Requirements: "Choose only: 'Yes' or 'No'. If there is a clear history of
smoking and drinking habits, answer 'Yes'; if the smoking and drinking
habits are denied, answer 'No'"
Document: "... Denies a history of smoking and drinking habits...."
Output: "row1: No"
'''
```

Merged Concepts

After merging all concepts, we filtered out those that appeared less than 5% of the time. A total of 117 concepts and terms were listed in Table S1 in Multimedia Appendix 2.

Question Template

Then we selected and transferred 68 concepts into question formats, for further Q&A scales. The detailed questions and their corresponding concepts are listed in Table S2 in Multimedia Appendix 2.

Scale Extraction

We identified that the optimal performance in Q&A scale extraction occurs with a concurrency of 3 requests, enhancing speed by 17.9% compared to a single request. Furthermore, we used a max_token restriction strategy, capping it at 20, to optimize inference speed.

Ultimately, within the 2D Q&A space formed by the answer scales, there were a total of 68 question columns and 25,709 observations (listed in Table 1).

We used accuracy and precision metrics for assessing the accuracy of LLMs across 1500 observations. Furthermore, we used 2 parameters—null ratio and time consumption—in 25,709 observations to evaluate the consistency and efficiency of the 2 LLMs, respectively.

Table 1. Question-and-answer scales for Qwen-14B-Chat (QWEN) and Baichuan2-13B-Chat (BAICHUAN).

Concepts	QWEN			BAICHUAN		
	Positive ratio (%)	Negative ratio (%)	Null ratio (%)	Positive ratio (%)	Negative ratio (%)	Null ratio (%)
Menstrual color	—[a]	—	0.00	—	—	0.67
Menstrual flow	—	—	0.00	—	—	0.80
Pregnancy weight gain[b]	—	—	1.60	—	—	2.53
Abdominal bloating	26.27	73.73	0.00	38.40	61.60	0.00
Abdominal pain	49.00	51.00	0.00	43.13	52.00	4.87
Amniocentesis	1.00	99.00	0.00	0.73	98.87	0.40
Aspirin use	1.40	98.60	0.00	1.40	98.60	0.00
Bilateral adnexal masses	4.87	95.13	0.00	0.73	99.20	0.07
Bilateral lower limb edema	51.53	48.47	0.00	4.87	95.13	0.00
Blood glucose screening	17.87	82.13	0.00	28.13	71.40	0.47
Cervical secretions	3.53	96.47	0.00	3.67	96.33	0.00
Chest tightness	2.87	97.13	0.00	4.40	95.53	0.07
Cold or flu	1.87	98.13	0.00	3.07	96.93	0.00
Convulsions	2.27	97.73	0.00	0.33	99.67	0.00
Dizziness	7.13	92.87	0.00	1.87	98.13	0.00
Drinking	0.00	100.00	0.00	0.00	100.00	0.00
Early pregnancy reaction or symptoms	85.93	14.07	0.00	47.53	52.13	0.33
Family history (asthma)	0.07	99.93	0.00	1.93	98.07	0.00
Family history (autoimmune disease)	0.20	99.80	0.00	3.07	96.93	0.00
Family history (diabetes mellitus)	1.67	98.33	0.00	2.33	97.67	0.00
Family history (drug allergy)	0.01	99.99	0.00	5.67	94.33	0.00
Family history (heart disease)	1.60	98.40	0.00	2.27	97.73	0.00
Family history (hematologic disease)	0.13	99.87	0.00	0.73	98.87	0.40
Family history (hypertension)	3.87	96.13	0.00	3.93	96.07	0.00
Family history (kidney disease)	0.07	99.93	0.00	1.40	98.53	0.07
Family history (mental illness)	0.07	99.93	0.00	0.47	99.47	0.07
Family history (neurological disease)	0.47	99.53	0.00	0.73	98.80	0.47
Family history (preeclampsia)	0.12	99.88	0.00	2.60	97.40	0.00
Family history (rheumatic disease)	0.07	99.93	0.00	1.47	98.53	0.00
Fetal paternal drinking history)	0.00	100.00	0.00	0.07	99.93	0.00
Fetal paternal history of genetic diseases	1.00	99.00	0.00	0.87	99.07	0.07
Fetal paternal smoking history	0.00	100.00	0.00	0.07	99.93	0.00
Fever	9.67	90.33	0.00	1.93	98.07	0.00
G6PD[c]	3.33	96.67	0.00	2.53	97.47	0.00
Headache	2.80	97.20	0.00	0.80	99.13	0.07
Insomnia	0.60	99.40	0.00	1.20	98.47	0.33
Mediterranean anemia screening	8.27	91.73	0.00	17.60	82.27	0.13
Palpitations	1.53	98.47	0.00	3.07	96.93	0.00
Personal history (antiphospholipid syndrome)	0.07	99.93	0.00	0.07	99.93	0.00
Personal history (chronic kidney disease)	0.80	99.20	0.00	1.07	98.93	0.00

Concepts	QWEN			BAICHUAN		
	Positive ratio (%)	Negative ratio (%)	Null ratio (%)	Positive ratio (%)	Negative ratio (%)	Null ratio (%)
Personal history (diabetes mellitus)	0.60	99.40	0.00	0.13	99.87	0.00
Personal history (drug allergy)	10.53	89.47	0.00	39.80	60.20	0.00
Personal history (dysmenorrhea)	24.40	75.60	0.00	21.20	78.80	0.00
Personal history (food allergy)	5.13	94.87	0.00	8.07	91.93	0.00
Personal history (heart disease)	1.67	98.33	0.00	0.47	99.53	0.00
Personal history (hematologic disease)	0.00	100.00	0.00	0.07	99.93	0.00
Personal history (hypertension)	7.40	92.60	0.00	0.93	99.07	0.00
Personal history (infectious disease)	1.93	98.07	0.00	3.80	96.20	0.00
Personal history (preeclampsia)	0.93	99.07	0.00	0.87	99.13	0.00
Personal history (surgery history)	35.67	64.33	0.00	36.27	63.67	0.07
Personal history (systemic lupus erythematosus)	0.20	99.80	0.00	0.20	99.80	0.00
Personal history (thalassemia)	1.00	99.00	0.00	0.80	99.13	0.07
Personal history (trauma history)	8.80	91.20	0.00	2.87	97.13	0.00
Personal history (viral hepatitis)	6.07	93.93	0.00	6.20	93.80	0.00
Poor pregnancy history (induced abortion)	0.07	99.93	0.00	0.13	99.87	0.00
Poor pregnancy history (miscarriage)	0.47	99.53	0.00	0.47	99.53	0.00
Poor pregnancy history (premature birth)	0.27	99.73	0.00	0.27	99.73	0.00
Prenatal screening	31.13	68.87	0.00	15.87	83.40	0.73
Regular prenatal check-ups	96.20	3.80	0.00	96.80	3.13	0.07
Smoking	0.07	99.93	0.00	0.07	99.93	0.00
Threatened abortion	6.20	93.80	0.00	5.80	94.00	0.20
Umbilical cord blood ratio	0.00	100.00	0.00	0.00	100.00	0.00
Use of antihypertensive drugs	1.73	98.27	0.00	2.13	97.87	0.00
Use of progestogen drugs	13.47	86.53	0.00	14.40	85.53	0.07
Vaginal bleeding	81.07	18.93	0.00	22.60	77.27	0.13
Vaginal discharge	33.00	67.00	0.00	48.47	51.53	0.00
Vaginal infection	25.27	74.73	0.00	16.20	82.73	1.07
Vaginal secretions	16.60	83.40	0.00	16.73	83.27	0.00

[a]Not applicable.

[b]The mean pregnancy weight gain was 13.73 (SD 24.12) for QWEN and 13.75 (SD 31.28) for BAICHUAN.

[c]G6PD: glucose-6-phosphate dehydrogenase.

Evaluation Metrics

Accuracy and Precision

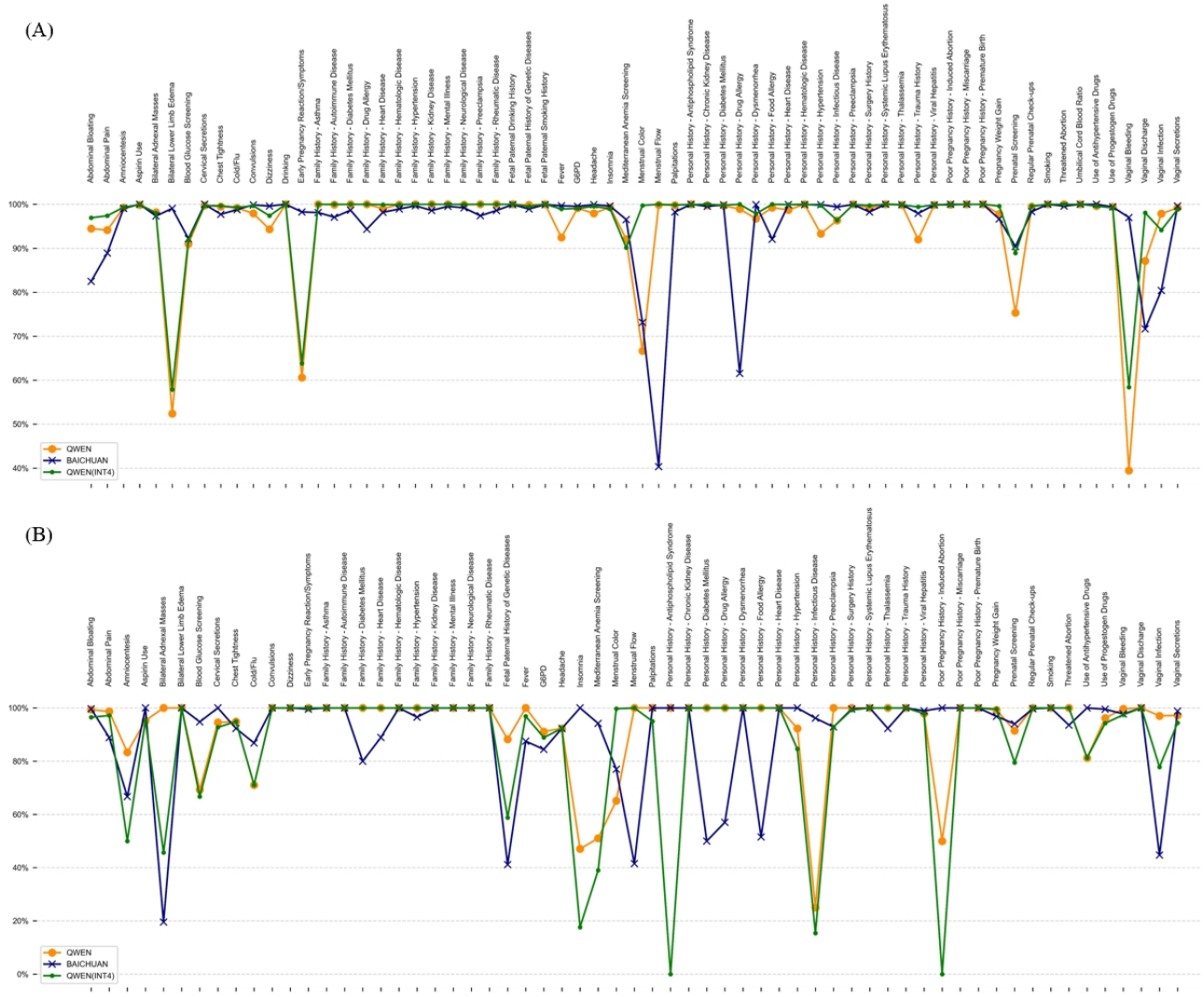
Figure 3A and 3B and Figure S1 (parts A and B) in Multimedia Appendix 2 illustrate the Q&A space for a sample chunk extracted by QWEN and BAICHUAN with the comparison of manual annotation. The figures demonstrate the exceptional accuracy and precision of QWEN and BAICHUAN. QWEN attained an average accuracy of 95.52% and an average precision of 92.93%, whereas BAICHUAN displayed an average accuracy of 95.86% and an average precision of 90.08%. These figures clearly indicate that the 2 LLMs have more concentrated errors in specific concepts, and overall, they achieve high levels of precision in most extractions.

LLMs demonstrated consistent performance across most questions and excelled in binary, well-defined medical history questions, often reaching 100% accuracy and precision. However, the accuracy performance varied significantly when dealing with questions that involved semantic ambiguities or definitional uncertainties. This inconsistency might be tied to the LLM's training and inference alignment. Notable disparities were observed in questions pertaining to menstrual color (QWEN: 1000/1500, 66.7%; BAICHUAN: 1097/1500, 73.1%), early pregnancy symptoms (QWEN: 909/1500, 60.7%; BAICHUAN: 1474/1500, 98.3%), vaginal bleeding (QWEN:

593/1500, 39.5%; BAICHUAN: 1455/1500, 97%), bilateral lower limb edema (QWEN: 786/1500, 52.4%; BAICHUAN: 1486/1500, 99.7%), and menstrual flow (QWEN: 1498/1500, 99.8%; BAICHUAN: 605/1500, 40.3%).

Apart from the above, the precision inconsistency performance of concepts could be attributed to their low true positive rate, like insomnia (QWEN: 3/17, 17.7%; BAICHUAN: 8/17, 47.1%), personal history—antiphospholipid syndrome (QWEN: 0/2, 0%; BAICHUAN: 1/2, 50%), and poor pregnancy history—induced abortion (QWEN: 1/2, 50%; BAICHUAN: 2/2, 100%). The exact precision is listed in Table S3 in Multimedia Appendix 2.

Figure 3. Accuracy and precision in the question-and-answer space. With the local expert's annotation of 1500 observations, parts A and B showcase a comparison of the accuracy and precision of QWEN, BAICHUAN, and QWEN(INT4) across various concepts. Our findings reveal that the performance trends of large language models are nearly uniform across different concepts in terms of accuracy while showing a discernible variation in precision. G6PD: glucose-6-phosphate dehydrogenase.



Null Ratio

As depicted in Table 1, both LLMs demonstrated superior performance with minimal null ratios. Specifically, QWEN (Figure S1A in Multimedia Appendix 2) exhibited a mean null ratio of 0.02%, in contrast to BAICHUAN (Figure S1B in Multimedia Appendix 2), which recorded a slightly higher null ratio of 0.21%. Failure of QWEN extraction was only in pregnancy weight gain (411/25,709, 1.60%), but failures of BAICHUAN extraction were mainly in symptoms (abdominal pain: 1252/25,707, 4.87%; vaginal infection: 275/25,709, 1.07%).

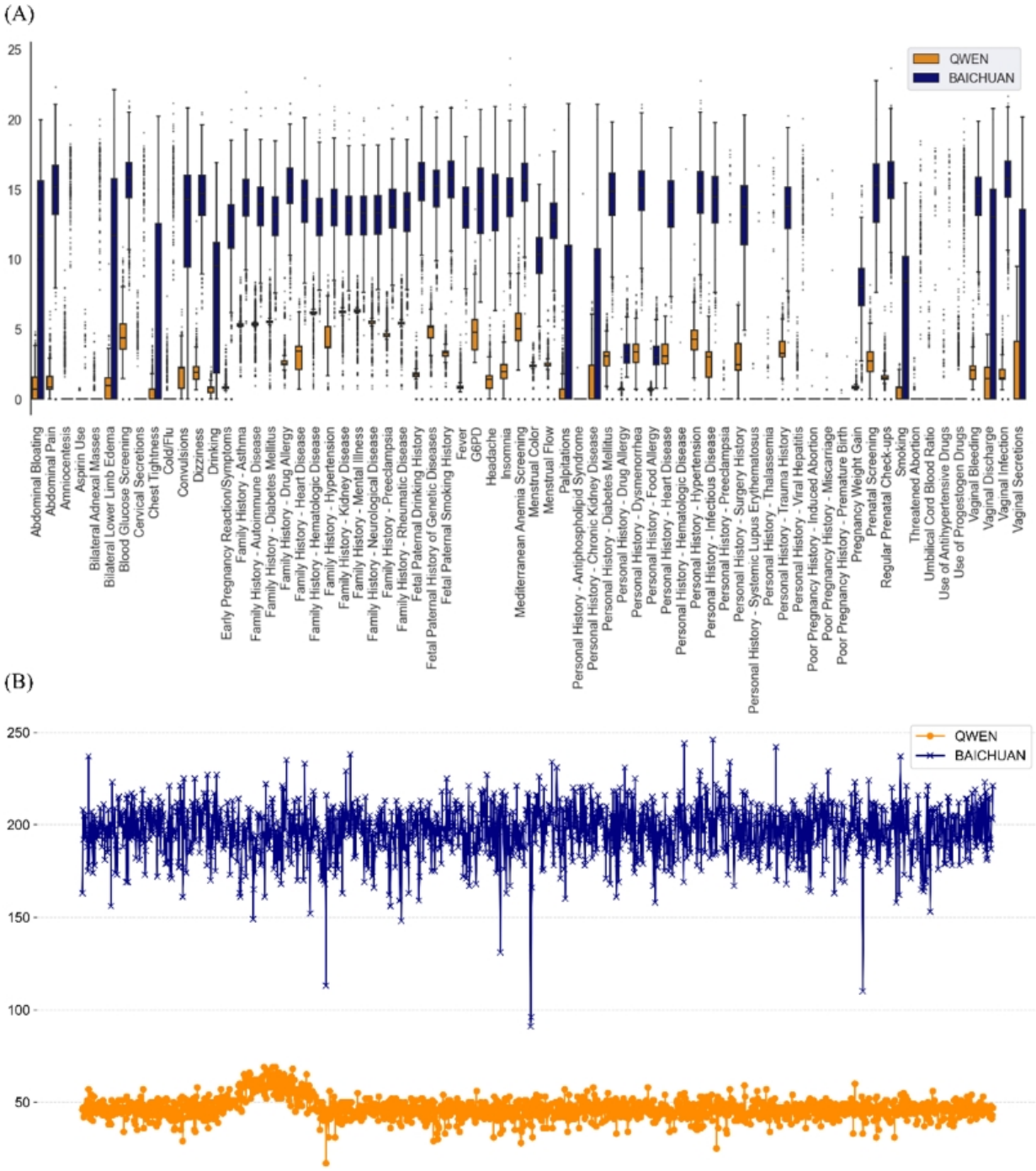
Time Consumption

We conducted a comparative analysis of the time performance between QWEN and BAICHUAN on various Q&A scales, discovering that BAICHUAN consistently exhibits higher time consumption across almost all scales, reaching up to 4 times that of QWEN, as illustrated in Figure 4B.

Figure 4A compares the time consumption of LLMs in extracting diverse concepts. Although there were significant differences across different concepts, overall, the LLMs demonstrated a consistent performance across these concepts. For queries with clear definitions and concise corpora, such as drug usage and previous pregnancy history, the time consumed was minimal. In the category of medical history, both models

exhibited uniform and stable performances (QWEN and BAICHUAN both revealed a time consumption ratio of 1:3).

Figure 4. Time consumption of question-and-answer (Q&A) scales for QWEN and BAICHUAN, measured in seconds. (A) Comparative distribution of time consumption for QWEN and BAICHUAN per concept, and QWEN exhibited notably lower extraction times across various Q&A scales compared to BAICHUAN. (B) Comparative distribution of the time consumption for 2 large language models per observation. G6PD: glucose-6-phosphate dehydrogenase.



Additional Research

In clinical practice, to address scenarios of resource constraints, we used a quantified version of the LLM in our study to validate the applicability of this approach. We used an official-release INT4 version of QWEN, QWEN(INT4). The model was deployed on an NVIDIA RTX 3090 GPU (24 GB).

With the same approach listed above, the performance of QWEN(INT4) achieved even better performance. Figure 3 and Figure S1C in Multimedia Appendix 2 demonstrate that the average accuracy of QWEN(INT4) is 97.28%, accompanied by a null ratio of 0%.

Despite a notable correlation in performance extraction between QWEN(INT4) and QWEN, QWEN(INT4) demonstrated superior efficiency on limited hardware, with an average of 31 seconds per observation, compared to 47 seconds for QWEN and 312 seconds for BAICHUAN.

Discussion

Principal Findings

In this study, the extracted scales incorporated not only the conventional features of interest but also less frequently mentioned dimensions in previous cohorts or guidelines. These included food and drug allergies (6.6% for food allergy and 25.2% for drug allergy), certain pregnancy symptoms (average positive ratio of 0.9% for insomnia and 2.3% for palpitations), menstrual conditions (22.8% for dysmenorrhea), medical history (1% for asthma family history and 0.27% for mental illness family history), and gestational intervention (13.93% for progestogen and 1.4% for aspirin).

As a naturally recruited cohort of pregnancy, the extracted features show comparable proportions or trends compared with similar studies, such as systemic lupus erythematosus (average positive ratio of 0.20% vs 0.03%-0.23% of similar cohorts [41,42]) and antiphospholipid syndrome (average positive ratio of 0.08% vs 0.02%-0.12% of similar cohorts [43]), thereby corroborating the accuracy of our approach.

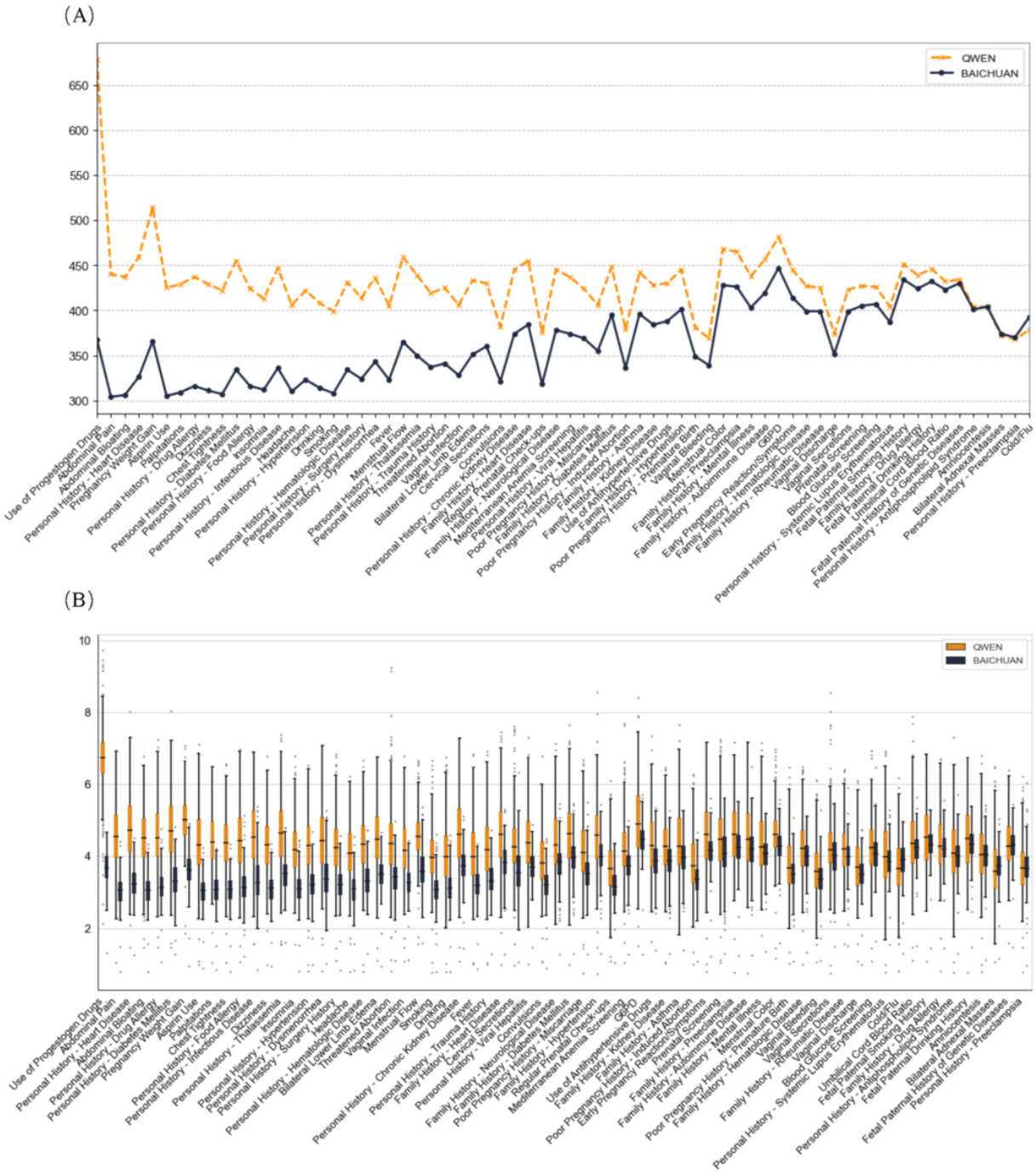
Additionally, certain scale deviations were revealed compared to similar studies, notably in fetal paternal smoking history (average positive ratio of 0.04% vs approximately 28.1%-40% in similar studies [44,45]). Although these deviations were few, we conducted a sample retracing to the original texts and determined that the extraction approach was not at fault and accurately reflected the original data. This discrepancy highlights persistent concerns [46] regarding the data quality in inpatient documentation, originating from patient self-reports and physician documentation, and vulnerability to recall and inquiry bias. Documentation varies among patients, influenced not only by patient conditions but also by physicians' writing habits. Thus, we regard our approach as a preexperimental data analysis. Despite the presence of biases or missing dimensions, the approach uncovers several dimensions absent in structured medical texts, and valuable insights could still be extracted from the data with appropriate statistics [47]. In clinical practice, preliminary interviews with documenting physicians are recommended prior to the selection of concepts to enhance data quality and mitigate potential biases.

In the context of the extraction process, even when deployed solely on a standard consumer-grade GPU (NVIDIA RTX 3090), the QWEN(INT4) completed the extraction of 25,709 observations and 68 features within 15 calendar days, averaging 48.9 seconds per observation. In practical applications, deploying 2 instances of QWEN(INT4) on a single graphics card, coupled with an additional deployment in CPU [36], is hypothesized to reduce the extraction to approximately 7 days. Furthermore, multi-GPU server clusters, prevalent in clinical environments, could markedly reduce processing times, potentially to the scale of hours.

In our study, we experimented with omitting the corpus extraction step, directly using the long text of each observation's chief complaints and medical histories as raw data for Q&A scale extraction. However, the experiment yielded poor performance in accuracy, precision, and time consumption, as illustrated in Figure 5 and Figure S2 in Multimedia Appendix 2.

These limitations appear significantly correlated with the current technological constraints of LLMs [32], which tend to generate "feature hallucinations" more frequently when processing extensive texts [48], leading to the loss of critical information. We believe that this issue will be resolved as the technology continues to evolve [49].

Figure 5. Approach performance when omitting corpus extraction step. (A) The average time consumed per question-and-answer (Q&A) interaction over 300 observations for both models. (B) Comparing the distribution of time consumption for QWEN and BAICHUAN in a single observation per Q&A scale. G6PD: glucose-6-phosphate dehydrogenase.



Limitations

In our experimental validation, we selected a limited set of concepts, comprising only 68 items, to balance the consideration of time constraints. Despite our efforts to encompass a broad scope, some dimensions inevitably remain unaddressed, which is a limitation in verifying efficiency and accuracy across all dimensions.

Furthermore, the raw data in this study was sourced exclusively from a single hospital, spanning nearly a decade. This duration, while significant, introduces limitations in the generalizability of our approach.

Additionally, the approach used only 2 LLMs. Although we anticipate that future LLMs will be compatible with the current approach, this assumption necessitates further experimental validation.

Conclusions

Our proposed approach offers a potential methodology for clinical text data analysis. It involves extracting and summarizing concepts from the comprehensive text of a defined population, thus selecting research directions of interest, and eventually generating analyzable features for the cohort. This approach demonstrates notable precision and could provide substantial data support for future research endeavors.

Acknowledgments

We are grateful to Yinyao Ma and the clinical team for their exceptional contributions to this project and we thank the technical support provided by China National GeneBank.

This work was supported by Guangxi Key Research and Development Program (AB22035056).

Data Availability

The data sets generated and analyzed during this study are not publicly available due to privacy or ethical restrictions but are available on request from the corresponding author.

Authors' Contributions

LW and YM contributed equally to this study. LW, HL, and YM participated in the study design and drafted the manuscript. YL and HL participated in data collection and outcome rule review. LW and WB performed the statistical analysis, and established machine learning models. YL helped to draft the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

A sample of chief complaints and medical histories.
[DOCX File , 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Additional statistics.
[DOCX File , 873 KB-Multimedia Appendix 2]

References

1. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Stats. Feb 14, 2021;13(6):1-19. [doi: 10.1002/wics.1549]
2. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng. 2018;2018:4302425. [doi: 10.1155/2018/4302425] [Medline: 29849998]
3. Varshini K, Uthra R. An approach to extract meaningful dData from unstructured clinical notes. In: Inventive Systems and Control. Singapore. Springer; Jun 08, 2021;581-590.
4. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. Mar 31, 2020;8(3):e17984. [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
5. Wu P, Yu A, Tsai C, Koh J, Kuo C, Chen ALP. Keyword extraction and structuralization of medical reports. Health Inf Sci Syst. Dec 03, 2020;8(1):18. [FREE Full text] [doi: 10.1007/s13755-020-00108-6] [Medline: 32269770]
6. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof. Nov 30, 2013;10:12. [FREE Full text] [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]
7. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. Am J Manag Care. Sep 2002;8(9):787-793. [FREE Full text] [Medline: 12234019]
8. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. Eval Health Prof. Sep 13, 2009;32(3):281-298. [doi: 10.1177/0163278709338561] [Medline: 19679636]
9. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health. Dec 22, 2022;1(12):e0000168. [FREE Full text] [doi: 10.1371/journal.pdig.0000168] [Medline: 36812634]
10. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform. Oct 2009;42(5):923-936. [FREE Full text] [doi: 10.1016/j.jbi.2009.07.007] [Medline: 19646551]
11. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020;21(140):1-67.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. [FREE Full text]
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

14. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform. Sep 12, 2019;7(3):e14830. [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
15. Wang S, Yilahun H, Hamdulla A. Biomedical named entity recognition based on MCBERT. 2022. Presented at: 2022 International Conference on Asian Language Processing (IALP); October 27-28, 2022; Singapore. [doi: 10.1109/ialp57159.2022.9961297]
16. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res. Jan 12, 2021;23(1):e19689. [FREE Full text] [doi: 10.2196/19689] [Medline: 33433395]
17. Zhang Z, Zhu L, Yu P. Multi-level representation learning for Chinese medical entity recognition: model development and validation. JMIR Med Inform. May 04, 2020;8(5):e17637. [FREE Full text] [doi: 10.2196/17637] [Medline: 32364514]
18. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. Jan 2018;77:34-49. [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
19. Polnaszek B, Gilmore-Bykovskyi A, Hovanes M, Roiland R, Ferguson P, Brown R, et al. Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction. Med Care. Oct 2016;54(10):e65-e72. [FREE Full text] [doi: 10.1097/MLR.0000000000000108] [Medline: 27624585]
20. Tarik AM, Sorin E, Symons J, Mayer E, Yaliraki S, Toni F. Extracting information from free text through unsupervised graph-based clustering: an application to patient incident records. arXiv. [FREE Full text] [doi: 10.48550/arXiv.1909.00183]
21. Choi HS, Song JY, Shin KH, Chang JH, Jang B. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J. Sep 2023;41(3):209-216. [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
22. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open. Oct 02, 2023;6(10):e2336997. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
23. Johnson S, King A, Warner E, Aneja S, Kann B, Bylund C. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr. Mar 01, 2023;7(2):a. [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
24. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. Mar 04, 2023;47(1):33. [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
25. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. Eur J Cardiovasc Nurs. Mar 12, 2024;23(2):122-126. [doi: 10.1093/eurjcn/zvad087] [Medline: 37603843]
26. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res. Dec 28, 2023;25:e51580. [FREE Full text] [doi: 10.2196/51580] [Medline: 38009003]
27. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng. Dec 16, 2023;51(12):2647-2651. [doi: 10.1007/s10439-023-03284-0] [Medline: 37328703]
28. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res. Sep 11, 2023;25:e49240. [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
29. Zhang L, Liu X, Li Z, Pan X, Dong P, Fan R. Dissecting the runtime performance of the training, fine-tuning, and inference of large language models. arXiv. [FREE Full text]
30. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J. A survey on large language model based autonomous agents. arXiv. ;01 [FREE Full text] [doi: 10.48550/arXiv.2308.11432]
31. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 28, 2022. [doi: 10.48550/arXiv.2201.11903]
32. Lei D, Li Y, Hu M, Wang M, Yun V, Ching E. Chain of natural language inference for reducing large language model ungrounded hallucinations. arXiv. [FREE Full text] [doi: 10.48550/arXiv.2310.03951]
33. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S. Self-consistency improves chain of thought reasoning in language models. arXiv. [FREE Full text]
34. ACOG committee opinion no. 743: low-dose aspirin use during pregnancy. Obstet Gynecol. Jul 2018;132(1):e44-e52. [doi: 10.1097/AOG.0000000000002708] [Medline: 29939940]
35. National IFHEG. National Institute for Health and Care Excellence (NICE). In: The Grants Register 2019. London, UK. Palgrave Macmillan; 2019;540.
36. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen Technical Report. arXiv. ;01 [FREE Full text] [doi: 10.48550/arXiv.2309.16609]

37. Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C. Baichuan 2: open large-scale language models. arXiv. [FREE Full text]
38. OpenAI. URL: https://openai.com/blog/openai-api [accessed 2024-03-12]
39. Zheng L, Chiang W, Sheng Y, Zhuang S, Wu Z, Zhuang Y. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv. ;01 [FREE Full text] [doi: 10.48550/arXiv.2306.05685]
40. SNOMED International. URL: http://www.ihtsdo.org/snomed-ct/ [accessed 2024-03-13]
41. Clowse ME, Jamison M, Myers E, James AH. A national study of the complications of lupus in pregnancy. Am J Obstet Gynecol. Aug 2008;199(2):127.e1-127.e6. [FREE Full text] [doi: 10.1016/j.ajog.2008.03.012] [Medline: 18456233]
42. Rees F, Doherty M, Grainge M, Lanyon P, Zhang W. The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. Rheumatology (Oxford). Nov 01, 2017;56(11):1945-1961. [doi: 10.1093/rheumatology/kex260] [Medline: 28968809]
43. Hwang J, Shin S, Kim Y, Oh Y, Lee S, Kim YH, et al. Epidemiology of antiphospholipid syndrome in Korea: a nationwide population-based study. J Korean Med Sci. Feb 10, 2020;35(5):e35. [FREE Full text] [doi: 10.3346/jkms.2020.35.e35] [Medline: 32030922]
44. Xu X, Rao Y, Wang L, Liu S, Guo JJ, Sharma M, et al. Smoking in pregnancy: a cross-sectional study in China. Tob Induc Dis. Jul 24, 2017;15(1):35. [FREE Full text] [doi: 10.1186/s12971-017-0140-0] [Medline: 28747859]
45. Yang Y, Liu F, Wang L, Li Q, Wang X, Chen JC, et al. Association of husband smoking with wife's hypertension status in over 5 million Chinese females aged 20 to 49 years. J Am Heart Assoc. Mar 20, 2017;6(3):e004924. [FREE Full text] [doi: 10.1161/JAHA.116.004924] [Medline: 28320748]
46. Leon N, Balakrishna Y, Hohlfeld A, Odendaal W, Schmidt B, Zweigenthal V, et al. Routine Health Information System (RHIS) improvements for strengthened health system management. Cochrane Database Syst Rev. Aug 13, 2020;8(8):CD012012. [FREE Full text] [doi: 10.1002/14651858.CD012012.pub2] [Medline: 32803893]
47. Miñarro-Giménez JA, Cornet R, Jaulent M, Dewenter H, Thun S, Gøeg KR, et al. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform. Mar 2019;123:37-48. [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.011] [Medline: 30654902]
48. Rawte V, Chakraborty S, Pathak A, Sarkar A, Towhidul ITS, Chadha A. The troubling emergence of hallucination in large language models -- an extensive definition, quantification, and prescriptive remediations. arXiv. ;01 [FREE Full text] [doi: 10.48550/arXiv.2310.04988]
49. Jones E, Palangi H, Simões C, Chandrasekaran V, Mukherjee S, Mitra A. Teaching language models to hallucinate less with synthetic tasks. arXiv. ;01 [FREE Full text] [doi: 10.48550/arXiv.2310.06827]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CoT: Chain of Thought
GPU: graphics processing unit
LLM: large language model
Q&A: question and answer
SNOMED CT: Systematised Nomenclature of Medicine Clinical Term

Edited by Q Jin; submitted 15.11.23; peer-reviewed by X Tannier, M Torii; comments to author 12.01.24; revised version received 23.01.24; accepted 14.02.24; published 29.03.24

Please cite as:
Wang L, Ma Y, Bi W, Lv H, Li Y
An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study
J Med Internet Res 2024;26:e54580
URL: https://www.jmir.org/2024/1/e54580
doi: 10.2196/54580
PMID: 38551633