

An Entity Extraction pipeline for Medical Text Records Utilizing Large Language Models: An Analytical Study

Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, Yuxiang Li

Submitted to: Journal of Medical Internet Research
on: November 15, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 42

Figures 43

Figure 1..... 44

Figure 2..... 45

Figure 3..... 46

Figure 4..... 47

Figure 5..... 48

Multimedia Appendixes 49

Multimedia Appendix 1..... 50

Multimedia Appendix 2..... 50

Multimedia Appendix 3..... 50

Multimedia Appendix 4..... 50

Multimedia Appendix 5..... 50

An Entity Extraction pipeline for Medical Text Records Utilizing Large Language Models: An Analytical Study

Lei Wang^{1, 2}; Yinyao Ma^{3*}; Wenshuai Bi¹; Hanlin Lv^{1*} MD; Yuxiang Li^{1, 2}

¹BGI Research, Wuhan Wuhan CN

²Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research, Shenzhen Shenzhen CN

³Department of Obstetrics, People's Hospital of Guangxi Zhuang Autonomous Region Nanning CN

*these authors contributed equally

Corresponding Author:

Hanlin Lv MD

BGI Research, Wuhan

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, Hubei, China

Wuhan

CN

Abstract

Background: Extracting valuable information from clinical text data is critical in disease progression studies. Traditional methods are often unable to cope with the complexity and volume of such data. The emergence of Large Language Models (LLMs) has opened new avenues, but they are challenged by critical issues such as data security and feature hallucination.

Objective: The primary objective of this study is to utilize a modular LLM approach to efficiently and accurately extract features from clinical text data, addressing the specific challenges of data security and feature hallucination, and improving upon the limitations of traditional methods.

Methods: In this study, we introduced a modular LLM approach to extract features from patient admission records. The process was divided into distinct steps: concept extraction, aggregation, question generation, corpus extraction, and Q&A scale extraction. Our method was evaluated on a dataset comprising 25,709 pregnancy cases from the People's Hospital of Guangxi Zhuang Autonomous Region, China, utilizing two low-parameter LLMs, Qwen-14B-Chat (QWEN) and Baichuan2-13B-Chat (BAICHUAN).

Results: The approach achieved high precision in features extraction, with QWEN and BAICHUAN showing average accuracies of 95.52% and 95.86%, respectively. The models demonstrated low null ratios (<0.21%) and varied time consumption. We also experimented the INT4-quantified version of QWEN (QWEN (INT4)) on a consumer-grade GPU, achieved even better performance (97.28% accuracy and a 0% null ratio).

Conclusions: This study demonstrates the effectiveness of a modular LLM approach in extracting clinical text data with high accuracy and efficiency. By breaking down the extraction process into manageable components, this approach offers a promising solution for textual features extraction from patient documentation.

(JMIR Preprints 15/11/2023:54580)

DOI: <https://doi.org/10.2196/preprints.54580>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/54580>, the full manuscript will be available to all users.



Original Manuscript

An Entity Extraction pipeline for Medical Text Records Utilizing Large Language Models: An Analytical Study

Lei Wang^{1,3†}, MS, Yinyao Ma^{2†}, MS, Wenshuai Bi¹, MS, Hanlin Lv^{1*}, MD, Yuxiang Li^{1,3*}, PhD

¹ BGI Research, Wuhan, 430074, China

² Department of Obstetrics, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, 530016, China

³ Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research, Shenzhen, 518083, China

[†]These authors contributed equally to this work.

Author contributions

L.W., Y.M. contributed equally to this study. L.W., H.L., Y.M. participated in study design and drafted the manuscript. X.L., H.L. participated in data collection, and outcome rule-review. L.W., W.B. performed the statistical analysis, established machine learning models. Y.L. helped to draft the manuscript. All authors read and approved the final manuscript.

*Corresponding Author:

Hanlin Lv

Tel: +86-18707190886

Postal Address: 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, Hubei, China

E-mail: 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 hotspot. 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.

Objectives: 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 Q&A scale extraction, which was tested via two low-parameter LLMs: Qwen-14B-Chat and Baichuan2-13B-Chat. A dataset of 25,709 pregnancy cases from the People's Hospital of Guangxi Zhuang Autonomous Region, China, was used for evaluation with help of local expert's annotation. The pipeline was evaluated with the metrics of accuracy and precision, null ratio and time consumption. Additionally, we also evaluated its performance via an INT4-quantified version of QWEN 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.

Keywords:

Clinical Data Extraction; Large Language Models; Feature Hallucination; Modular Approach; Unstructured Data Processing.

Introduction

Clinical text data has been widely recognized in data research due to its encompassment of multi-source[1, 2] (including patient subjective statements, past objective facts, and doctor's diagnostic processes and summary records). Extracting useful information from text data could serve as a crucial supplement to the study of disease progression and complement objective indicators dependent on laboratory tests and examinations[3], which has consistently been a hot research topic[4, 5].

Historically, method for text data extraction mainly includes: 1) Manual annotation: firstly, scales are designed based on clinical and research experience, followed by manual field extraction[6-8]; 2) Rule Extraction: utilizing concepts from established knowledge base (such as International Classification of Diseases, Tenth Revision (ICD-10))[9] for concept term extraction, typically basing on similarity algorithms and manual assistance to extract term and its attributes (negations, dependency relationships, etc.); [10]; 3) NER/NLP algorithms: utilizing supervised learning (pre-trained models like T5[11], BERT[12], and its 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]: 1) it is challenging to determine the dimension of extracted features initially, and from another perspective, confining the feature dimension means constraining research scope from the outset[19]; 2) given the inherent subjectivity and potential biases of recording subject, solely relying on algorithms without annotation typically results in low accuracy and recall[20]; 3) achieving higher accuracy with a broader feature scope, and the required human effort involved, is typically non-linear [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 nature 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, numerous considerations and limitations in application: 1) High-performing LLMs like OpenAI ChatGPT and Google Bard[25, 26], currently are not open source, and patient data need to be submitted to their platform for analysis, presenting security challenges[27, 28]; 2) Open-source LLMs performing high intelligence generally require large parameters number (10~100B), which are hard to support in consumer-level GPUs[29]; 3) Low-parameter (around 10B) LLMs, typically require multi-strategy support when dealing with tasks in certain vertically segmentation[30] (like fine-tuning, knowledge base/knowledge graph support, complex Chain of Thought (CoT)[31] and its derivatives, and even global training etc.), and are accompanied by various anomaly issues, including feature hallucination.

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

In this study, we aim to extract valuable features from a series of given patient admission records, which encompass the chief complaint and the medical histories. In light of this task, we introduce a modular LLMs approach, which divides the entire extraction path into several smaller steps and each modular LLMs handle these basic steps automatically. We adopted the core idea of LLM agents[30] and Self-consistency with CoT (CoT-SC)[33].

To experiment the approach, we implemented two low-parameter LLM 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 the 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 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 (ACOG)[36] and the 2019 guidelines from the National Institute for Health and Care Excellence (NICE)[37].

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

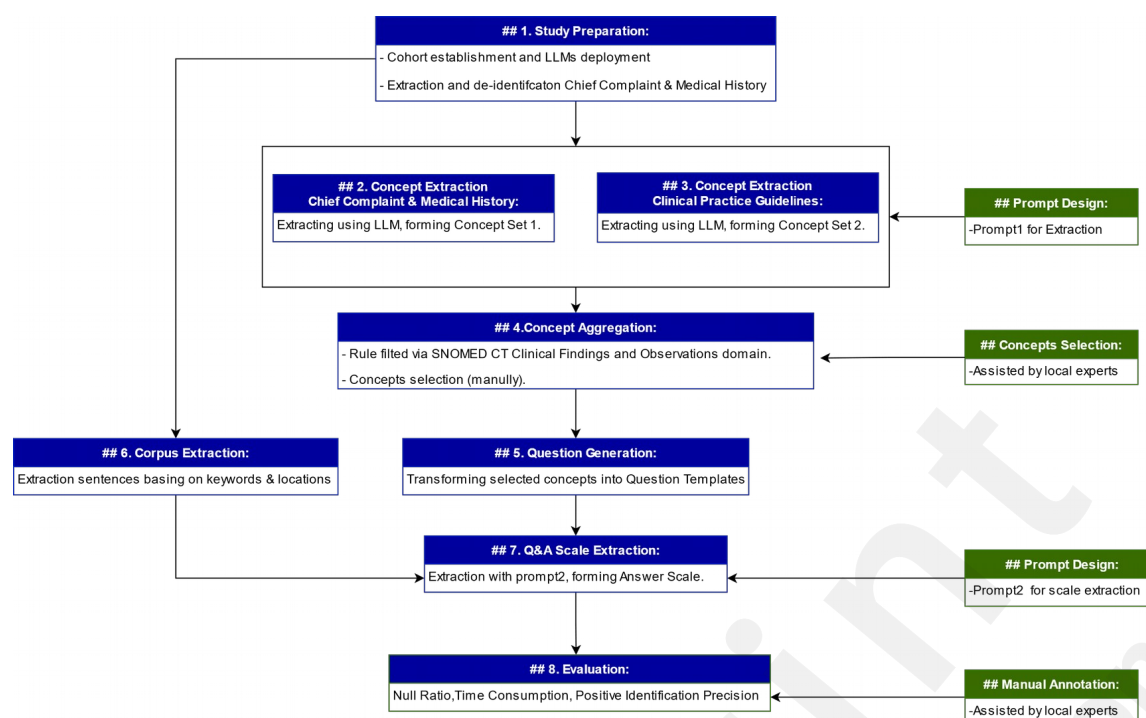
Model Deployment

We deployed 2 most exemplary LLMs in Chinese domain till Sep. 2023 in an intranet security environment: Qwen-14B-Chat(QWEN)[34] and Baichuan2-13B-Chat(BAICHUAN)[35]. Through independent deployment. In the environment, the server cluster utilized NVIDIA DGX-A100 (2×40G) GPU nodes. The QWEN used 29GB of storage and 27GB of GPU memory, while the BAICHUAN used 26GB of storage and 28.9GB 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 introduce an approach that autonomously extracts valuable textual features. Diverging from traditional LLM applications, we employed an “external-COT” strategy, dividing the process into several controllable steps, as illustrated in Figure 1.

Figure 1. Flow chart of Extraction approach. The extraction approach could be divided into 4 parts: 1) concept preparation: extracting existing concepts from corpus, and selecting concerning concepts; 2) corpus preparation: de-identifying raw data and preparing corpus in accordance with selected concept; 3) prompt design: for different LLM tasks; 4) Q&A scales: transforming concepts into question templates and extracting corresponding scales by LLMs.



Prompt design

The design of prompt templates is fundamental to efficient and accurate extraction. Prior to processing the entire dataset, 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 four-paragraph structure, referring to the prompt engineering suggestions of QWEN and BAICHUAN: 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, employs the Chain of Thoughts (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 utilizing 100 observations at different levels of concurrency to select the most optimal configuration.

Concept extraction & Aggregation

We initially extracted all discernible concepts from chief complaints and medical histories using LLMs with a designed Prompt1, 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 ACOG 2018 and NICE 2019.

As we defined in Prompt1, the extracted concepts were set in format of 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 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 examinations). 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 following 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 (e.g., confusing current medical history with past medical history, or confusing the patient's own medical history with that of family members), we pre-extracted 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 labelled with the corresponding question templates for the subsequent extraction of question-and-answer (Q&A) scales.

The refined corpus, combined with corresponding question templates, guided a systematic extraction process with two LLMs, and formed 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 three-fold assessment strategy was developed, as explained below:

Accuracy and precision

A subset of 1,500 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 (symbols and gibberish) outputs were identified as null output 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 two 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 (Ref. No. KT-KJT-2021-67). The requirement for informed consent was waived, due to the retrospective nature of the study, and all clinical data were de-identified and anonymized.

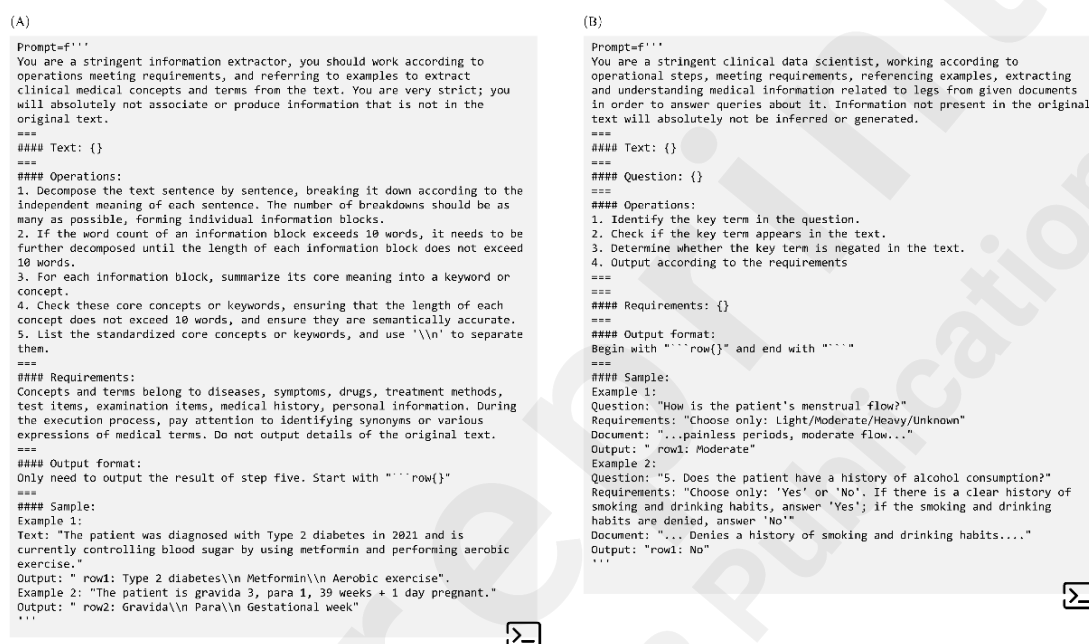
Results

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 Q&A scales. A four-paragraph structure was implemented for prompt design utilizing few-shot CoT. The original version, written in Chinese, was translated into English. (A) Extracting concepts and terms, and (B) Scaling questions. As both deployed LLMs were pre-trained with ICD-10 corpus, they could directly engage in concept extraction.



Merged concepts

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

Table S1 (Multimedia Appendix 2). Merged concept sets and classification.

Class	Concepts from chief complaints and medical histories	Concepts from clinical practice guidelines
Basic Information	Occupation, Age, Education (status), Marital Status, Date of admission to hospital, Date of last menstrual period, Estimated date of delivery, Pregnancy status findings, Finding of number of pregnancies, Number of fetuses in utero, Menstrual cycle finding, Menstrual blood loss finding, Color of menstrual blood	Occupation, Age, Ethnic Group, Marital Status, Gestational age, Finding of number of pregnancies, Number of fetuses in utero, History of tobacco use, History of alcohol consumption behavior
Condition/State Present	Nausea, Vomiting, Anorexia, Sense of fetal movement, Pyrexia, Feeling of chest pressure, Palpitations, Abdominal pain, Abdominal bloating and distension,	Assisted reproductive technology, Threat of spontaneous abortion, Insomnia, Early

	Common cold, Vaginal discharge, Bleeding, Dizziness, Headache, Convulsion, Edema of both lower legs, Antenatal screening, Mass of both adnexa, Vaginal infection, Vertex presentation of fetus	pregnancy symptoms,
Indicators Requiring Specific Values	Cord around neck, Amniotic fluid index, Diameter of fetal biparietal, Thickness of placenta, Grading of placental function	Change in weight, Length of fetal femur
Pregnancy Indicator Results	Finding of amniotic fluid, Non-invasive prenatal screening, Finding of chromosomal analysis of amniotic fluid, Result of Ureaplasma urealyticum test, Ultrasonography (scan), Three-dimensional color doppler sonography, Screening for glucose tolerance, Screening for thalassemia, Glucose-6-phosphate dehydrogenase measurement, Triplicate immunization status, Quantitation of Hepatitis B test, Coagulation test, Eugenic screening, Thyroid function study, Finding of cervical secretions, Screening for Down's syndrome, Vaginal swab culture	Three-dimensional color doppler sonography, Finding of fetal umbilical blood Doppler
Pregnancy Treatment Interventions	Progesterone, Dydrogesterone, Folic acid, Azithromycin, Aspirin, Methyldopa, Labetalol	Aspirin, Labetalol
History or Present Illness Description	Dysmenorrhea, Previous pre-eclampsia, Diabetes mellitus, Blood disorder, Heart disease, Thalassemia, Viral hepatitis, Chronic kidney disease, Tuberculosis, Injury, History of surgical procedure, Sexually transmitted infectious disease, Genetic disease, History of infectious disease, Food allergy, Drug allergy	Hypertension, Previous preterm delivery, Systemic lupus erythematosus, Antiphospholipid syndrome, History of spontaneous abortion, History of deliberate termination of pregnancy
Family History	Family history of cardiac disorder, Family history of diabetes mellitus, Family history of hematological disease, Family history of liver disease, Family history of kidney disease, Family history of mental health disorder, Family history of neurological disorder, Family history of rheumatologic disorder, Family history of endocrine disease, Family history of autoimmune disorder, Family history of allergy, Family history of	Family history of hypertension Family history of preeclampsia

sudden death, Family history of obesity, Family history of asthma, Family history of fracture, Family history of surgical procedure, Family history of adverse reaction to drug

Health Status Hereditary Disease, Present Illness
of the Fetus History, Smoking History, Drinking
Father History Smoking History, Drinking History

Question template

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

Table S2 (Multimedia Appendix 2). Concept and question list

	Concept	Question
1	Abdominal Bloating	Did the patient experience abdominal bloating during pregnancy? Provide a reason.
2	Abdominal Pain	Did the patient experience abdominal pain during pregnancy? Provide a reason.
3	Amniocentesis	Was the patient's amniocentesis result abnormal during pregnancy? Provide a reason.
4	Aspirin Use	Has the patient ever used aspirin? Provide a reason.
5	Bilateral Adnexal Masses	Did the patient have an adnexal mass found during pregnancy? Provide a reason.
6	Bilateral Lower Limb Edema	Did the patient have edema in the lower extremities during pregnancy? Provide a reason.
7	Blood Glucose Screening	Is the patient's blood glucose screening or OGTT test result abnormal? Provide a reason.
8	Cervical Secretions	Is the patient's cervical secretion test result abnormal? Provide a reason.
9	Chest Tightness	Did the patient experience chest tightness during pregnancy? Provide a reason.
10	Cold/Flu	Did the patient catch a cold during pregnancy? Provide a reason.
11	Convulsions	Did the patient have convulsions during pregnancy? Provide a reason.
12	Dizziness	Did the patient experience dizziness during pregnancy? Provide a reason.
13	Drinking	Does the patient report a history of alcohol consumption? Provide a reason.
14	Early Pregnancy Reaction/Symptoms	Did the patient experience nausea, vomiting, anorexia, or other early pregnancy reactions? Provide a reason.
15	Family History - Asthma	Is there a history of asthma in the patient's family? Provide a reason.
16	Family History - Autoimmune Disease	Is there a history of autoimmune diseases in the patient's family? Provide a reason.
17	Family History - Diabetes Mellitus	Is there a history of diabetes in the patient's family? Provide a reason.
18	Family History -	Is there an allergy history in the patient's family?

- Drug Allergy Provide a reason.
- 19 Family History - Is there a history of heart disease in the patient's family? Provide a reason.
- 20 Family History - Is there a history of blood disorders in the patient's Hematologic Disease family? Provide a reason.
- 21 Family History - Is there a history of hypertension in the patient's Hypertension family? Provide a reason.
- 22 Family History - Is there a history of kidney disease in the patient's Kidney Disease family? Provide a reason.
- 23 Family History - Is there a history of mental illness in the patient's Mental Illness family? Provide a reason.
- 24 Family History - Is there a history of neurological disease in the Neurological Disease patient's family? Provide a reason.
- 25 Family History - Has there been a history of pre-eclampsia in the Preeclampsia patient's family? Provide a reason.
- 26 Family History - Is there a history of rheumatic disease in the patient's Rheumatic Disease family? Provide a reason.
- 27 Fetal Paternal Drinking History Does the physiological father of the fetus have a history or habit of drinking alcohol? Provide a reason.
- 28 Fetal Paternal History of Genetic Diseases Does the physiological father of the fetus have a history of genetic diseases? Provide a reason.
- 29 Fetal Paternal Smoking History Does the physiological father of the fetus have a history or habit of smoking? Provide a reason.
- 30 Fever Did the patient have a fever during pregnancy? Provide a reason.
- 31 G6PD Is the patient's glucose-6-phosphate dehydrogenase (G6PD) level abnormal? Provide a reason.
- 32 Headache Did the patient experience headaches during pregnancy? Provide a reason.
- 33 Insomnia Did the patient suffer from insomnia during pregnancy? Provide a reason.
- 34 Mediterranean Anemia Screening Is the patient's thalassemia screening result abnormal? Provide a reason.
- 35 Menstrual Color What is the color of the patient's menstruation? (Choose from: light red/red/dark red/other/unknown)
- 36 Menstrual Flow What is the patient's menstrual volume? (Choose from: light/medium/heavy/unknown)
- 37 Palpitations Did the patient experience palpitations during pregnancy? Provide a reason.
- 38 Personal History - Does the patient have a history of antiphospholipid Antiphospholipid Syndrome syndrome? Provide a reason.
- 39 Personal History - Does the patient have a history of chronic kidney Chronic Kidney Disease disease? Provide a reason.
- 40 Personal History - Does the patient have a history of diabetes? Provide a Diabetes Mellitus reason.
- 41 Personal History - Does the patient admit to allergies to certain Drug Allergy medications? Provide a reason.

- 42 Personal History - Does the patient experience dysmenorrhea? Provide a reason.
Dysmenorrhea
- 43 Personal History - Does the patient admit to allergies to certain foods? Provide a reason.
Food Allergy
- 44 Personal History - Does the patient have a history of heart disease? Provide a reason.
Heart Disease
- 45 Personal History - Does the patient have a history of blood disorders? Provide a reason.
Hematologic Disease
- 46 Personal History - Does the patient have a history of hypertension? Provide a reason.
Hypertension
- 47 Personal History - Does the patient admit to a history of infectious diseases (such as tuberculosis, gonorrhea, syphilis, etc.)? Provide a reason.
Infectious Disease
- 48 Personal History - Does the patient have a history of pre-eclampsia? Provide a reason.
Preeclampsia
- 49 Personal History - Does the patient have a history of surgery? Provide a reason.
Surgery History
- 50 Personal History - Does the patient have a history of systemic lupus erythematosus? Provide a reason.
Systemic Lupus Erythematosus
- 51 Personal History - Does the patient have a history of thalassemia? Provide a reason.
Thalassemia
- 52 Personal History - Does the patient have a history of trauma? Provide a reason.
Trauma History
- 53 Personal History - Does the patient have a history of viral hepatitis? Provide a reason.
Viral Hepatitis
- 54 Poor Pregnancy History - Does the patient admit to a history of induced abortion? Provide a reason.
Induced Abortion
- 55 Poor Pregnancy History - Does the patient admit to a history of miscarriage? Provide a reason.
Miscarriage
- 56 Poor Pregnancy History - Does the patient admit to a history of preterm birth? Provide a reason.
Premature Birth
- 57 Pregnancy Weight Gain - What is the patient's pregnancy weight change? (Answer with a number only.)
- 58 Prenatal Screening - Was the patient's Down syndrome screening or non-invasive DNA test result abnormal during pregnancy? Provide a reason.
- 59 Regular Prenatal Check-ups - Did the patient undergo regular prenatal check-ups during pregnancy? Provide a reason.
- 60 Smoking - Does the patient report a personal history of smoking? Provide a reason.
- 61 Threatened Abortion - Does the patient have a history of threatened abortion? Provide a reason.
- 62 Umbilical Cord Blood Ratio - Is the fetal umbilical cord blood ratio result of the patient abnormal? Provide a reason.
- 63 Use of Antihypertensive Drugs - Has the patient ever used antihypertensive medication? Provide a reason.

64	Use of Progestogen Drugs	Has the patient ever used progestogen medication? Provide a reason.
65	Vaginal Bleeding	Did the patient experience vaginal bleeding during pregnancy? Provide a reason.
66	Vaginal Discharge	Did the patient experience vaginal discharge during pregnancy? Provide a reason.
67	Vaginal Infection	Did the patient experience vaginal infections during pregnancy? Provide a reason.
68	Vaginal Secretions	Is the patient's vaginal secretion test result abnormal? Provide a reason.

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 employed a max_token restriction strategy, capping it at 20, to optimize inference speed.

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

Table 1. Q&A scales for QWEN and BAICHUAN

Concepts	QWEN				BAICHUAN			
	Positive	Negative	Null Ratio	Mean (SD)	Positive	Negative	Null Ratio	Mean (SD)
Menstrual Color	-	-	0.00%	-	-	-	0.67%	-
Menstrual Flow	-	-	0.00%	-	-	-	0.80%	-
Pregnancy Weight Gain	-	-	1.60%	13.73 (24.12)	-	-	2.53%	13.75 (31.28)
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/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/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	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%	-
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	0.27%	99.73%	0.00%	-	0.27%	99.73%	0.00%	-

History - Premature Birth			%				%	
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 %	-

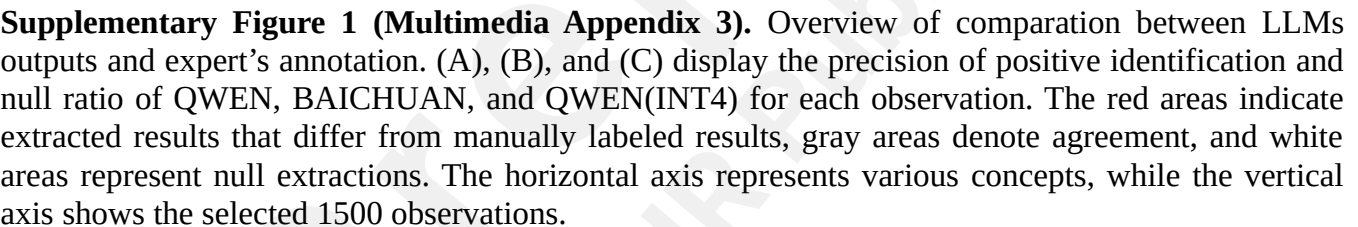
We employed accuracy and precision metrics for assessing the accuracy of LLMs across 1,500 observations. Furthermore, we utilized two parameters—null ratio and time consumption—in 25,709 observations to evaluate the consistency and efficiency of two LLMs, respectively.

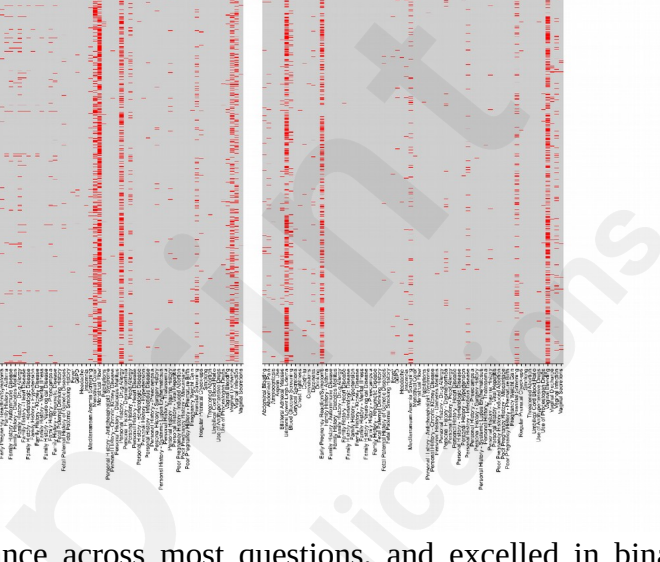
Evaluation metrics

Accuracy and precision

Figure 3 (A), Figure 3 (B) and Supplementary Figure 1 (A) and (B) 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 two LLMs have more concentrated errors in specific concepts, and overall, they achieve high levels of precision in most extractions.

Figure 3. Accuracy and precision in Q&A space. With the local expert's annotation of 1500-observation, (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 LLMs are nearly uniform across different concepts in terms of accuracy, while showing a discernible variation in precision.





Apart from the above, the precision inconsistency performance were as their low true positive rate, like Insomnia (QWEN 17.7%, 3 out of 17; BAICHUAN 47.1%, 8 out of 17), Personal History - Antiphospholipid Syndrome(QWEN 0%, 0 out of 2; BAICHUAN 50%, 1 out of 2), Poor Pregnancy History - Induced Abortion(QWEN 50%, 1 out of 2; BAICHUAN 100%, 2 out of 2). The exact precision is listed in Table S3.

Concept	QWEN	BAICHUAN	QWEN	True
		N	INT4	positive
Abdominal Bloating	313	314	304	315
Abdominal Pain	656	591	646	665
Amniocentesis	5	4	3	6
Aspirin Use	20	21	20	21
Bilateral Adnexal Masses	46	9	21	46
Bilateral Lower Limb Edema	59	59	59	59
Blood Glucose Screening	239	327	230	345

Cervical Secretions	52	55	51	55
Chest Tightness	37	36	37	39
Cold/Flu	27	33	27	38
Convulsions	3	3	3	3
Dizziness	22	22	22	22
Early Pregnancy	698	695	698	698
Reaction/Symptoms				
Family History - Asthma	1	1	1	1
Family History - Autoimmune Disease	2	2	2	2
Family History - Diabetes Mellitus	25	20	25	25
Family History - Heart Disease	9	8	9	9
Family History - Hematologic Disease	1	1	1	1
Family History - Hypertension	58	56	58	58
Family History - Kidney Disease	1	1	1	1
Family History - Mental Illness	1	1	1	1
Family History - Neurological Disease	6	6	6	6
Family History - Rheumatic Disease	1	1	1	1
Fetal Paternal History of Genetic Diseases	15	7	10	17
Fever	32	28	31	32
G6PD	41	38	40	45
Headache	12	12	12	13
Insomnia	8	17	3	17
Mediterranean Anemia Screening	123	227	94	241
Menstrual Color	920	1087	1408	1412
Menstrual Flow	1428	593	1428	1428
Palpitations	20	20	19	20
Personal History - Antiphospholipid Syndrome	1	1	0	1
Personal History - Chronic Kidney Disease	9	9	9	9
Personal History - Diabetes Mellitus	4	2	4	4
Personal History - Drug Allergy	142	81	142	142
Personal History - Dysmenorrhea	317	317	317	317
Personal History - Food Allergy	66	34	66	66

Personal History - Heart Disease	6	6	6	6
Personal History - Hypertension	12	13	11	13
Personal History - Infectious Disease	13	50	8	52
Personal History - Preeclampsia	14	13	13	14
Personal History - Surgery History	523	521	519	523
Personal History - Systemic Lupus Erythematosus	3	3	3	3
Personal History - Thalassemia	13	12	13	13
Personal History - Trauma History	12	12	12	12
Personal History - Viral Hepatitis	91	92	91	93
Poor Pregnancy History - Induced Abortion	1	2	0	2
Poor Pregnancy History - Miscarriage	7	7	7	7
Poor Pregnancy History - Premature Birth	4	4	4	4
Pregnancy Weight Gain	1441	1406	1446	1452
Prenatal Screening	107	110	93	117
Regular Prenatal Check-ups	1435	1431	1435	1435
Smoking	1	1	1	1
Threatened Abortion	93	87	93	93
Use of Antihypertensive Drugs	26	32	26	32
Use of Progestogen Drugs	201	208	197	209
Vaginal Bleeding	309	303	302	310
Vaginal Discharge	302	302	302	302
Vaginal Infection	358	165	287	369
Vaginal Secretions	244	248	237	251

Null ratio

As depicted in Table 1, both LLMs demonstrated superior performance with minimal null ratios. Specifically, QWEN (refer to Supplementary Figure 1 (A)) exhibited a mean null ratio of 0.02%, in contrast to BAICHUAN (see Supplementary Figure 1 (B)), which recorded a slightly higher null ratio of 0.21%. Failure of QWEN extraction was only in Pregnancy Weight Gain (1.60%, 411 out of 25709), but failures of BAICHUAN extraction were mainly in symptoms (Abdominal Pain, 4.87%, 1252 out of 25707; Vaginal Infection, 1.07%, 275 out of 25709; etc.)

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 4(B).

Figure 4. Time Consumption of Q&A scales for QWEN and BAICHUAN, measured in seconds. Figure (A) presents 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. and figure (B) depicts the comparative distribution of the time consumption for two LLMs per observation.

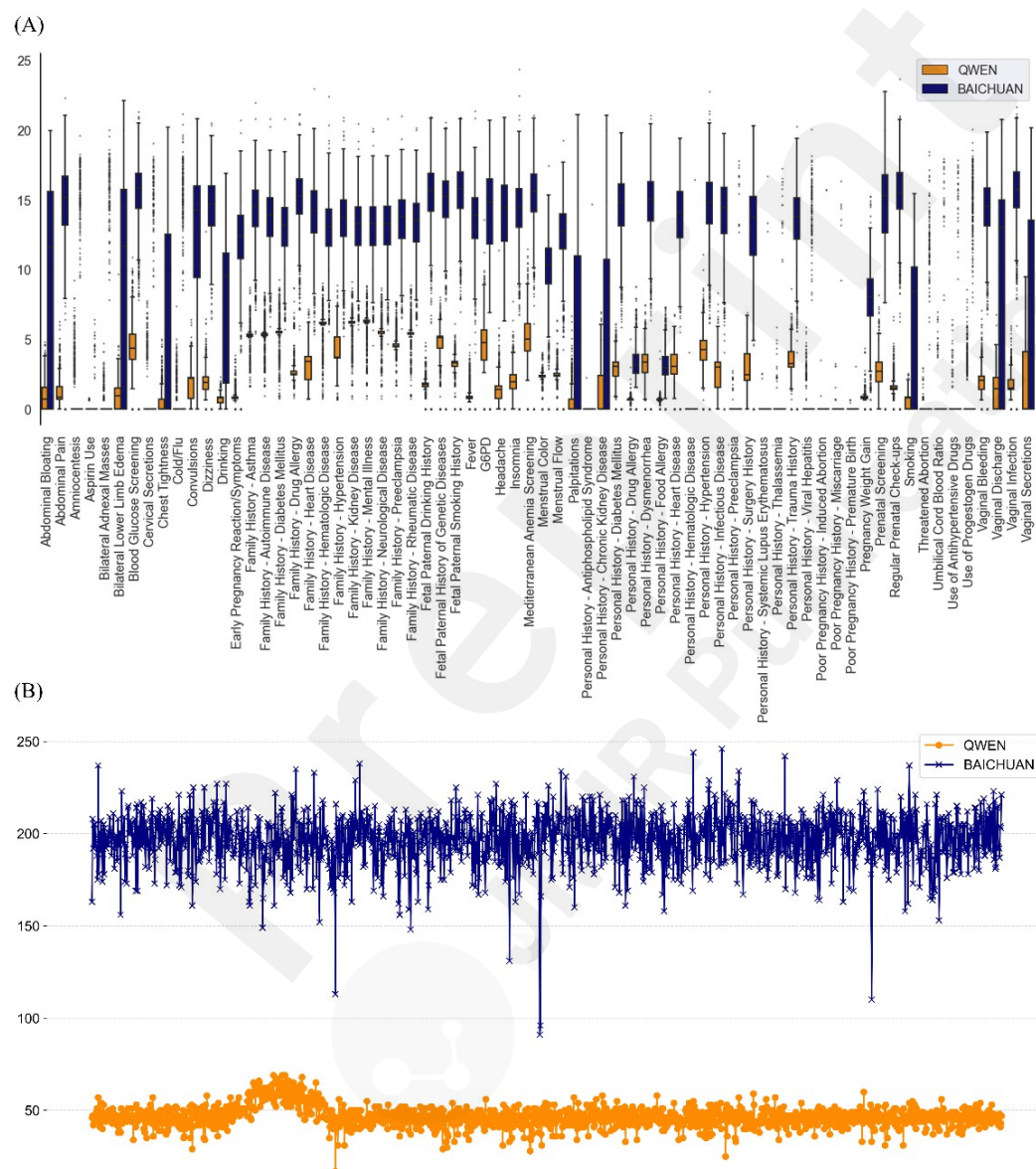


Figure 4(A) compares the time consumption of LLMs in extracting diverse concepts. Although there were significant differences across different concepts, overall, the LLMs demonstrated a consistent performance across these concepts. For queries with clear definitions and concise corpora, such as drug usage and previous pregnancy history, the time consumed was minimal. In the category of medical history, both models exhibit uniform and stable performance (QWEN and BAICHUAN, revealing a time consumption ratio of 1:3).

Additional research

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

With the same approach listed above, the performance of QWEN(INT4) achieved even better performance. Figure 3 and Supplementary Figure 1 (C) 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 in limited hardware, with an average of 31 seconds per observation, compared to 47 seconds for QWEN and 312 seconds for BAICHUAN.

Discussion

Principle findings

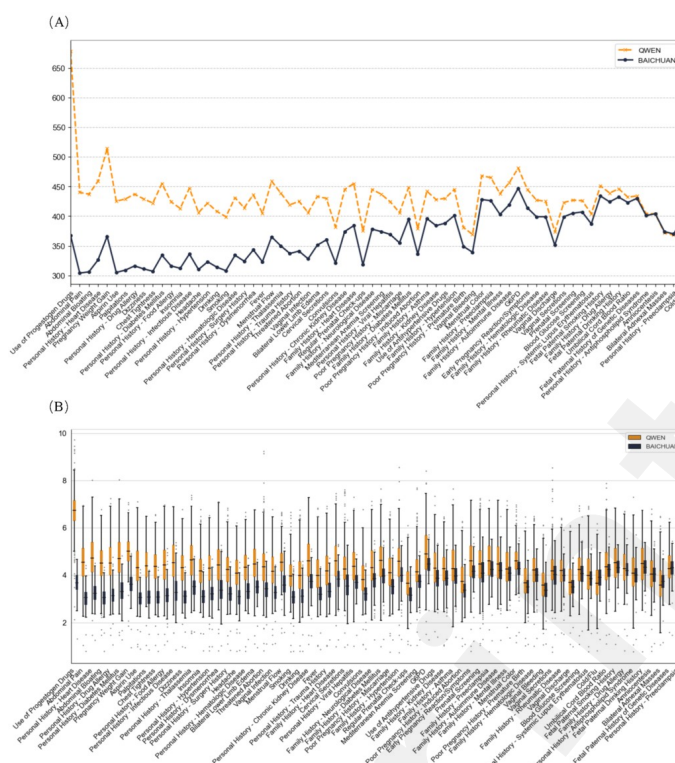
In this study, the extracted scales incorporated not only the conventional features of interest but also less frequently mentioned dimensions in previous cohorts or guidelines. These included food and drug allergies (6.6% for food allergy, 25.2% for drug allergy), certain pregnancy symptoms (average positive ratio of 0.9% for insomnia, 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 progesterone, 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%, versus 0.03%~0.23% of similar cohorts[41, 42]), and antiphospholipid syndrome (average positive ratio of 0.08%, versus 0.02%~0.12% of similar cohorts[43]) thereby corroborating the accuracy of our approach.

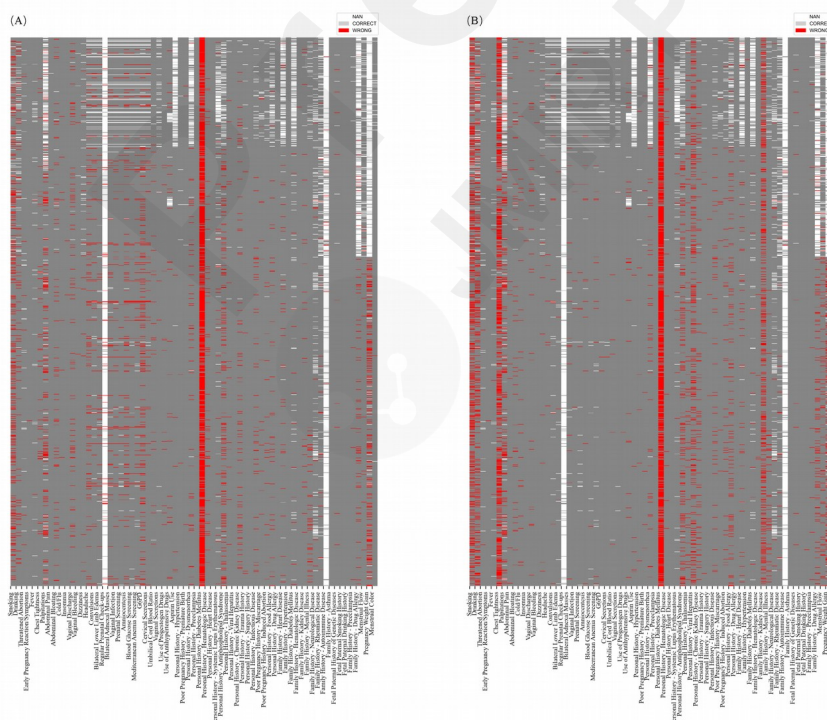
Additionally, certain scale deviations were revealed compared to similar studies, notably in Fetal Paternal Smoking History (0.04%, approximately 28.1%~40% in similar studies[44, 45]). Although these deviations were few, we conducted a sample retracing to the original texts and determined that the extraction 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 vulnerable 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 pre-experimental 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 observational and 68 features within 15 calendar days, averaging 48.9 seconds per observation. In practical applications, deploying two instances of QWEN(INT4) on a single graphics card, coupled with an additional deployment in CPU[34], 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 utilizing 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 Supplementary Figure 2.

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



Supplementary Figure 2 (Multimedia Appendix 4). Overview of comparison between LLMs outputs and expert's annotation when omitting corpus extraction step. Figure (A) and Figure (B) depict, respectively, the positive identification accuracy and null ratio of QWEN and BAICHUAN over 300 manually annotated observations, with each row representing a single observation. These figures highlight significant deficiencies in extracting medical histories and specific symptoms, contributing to an average positive identification accuracy of less than 70%.



These limitations appear significantly correlated with the current technological constraints of LLM[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].

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 to verify 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 employed only two LLMs. While we anticipate that future LLMs will be compatible of current approach, this assumption necessitates further experimental validation.

Conclusion

Our proposed approach offers a potential methodology for clinical text data analysis. It involves extracting and summarizing concepts from 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.

Source of Funding

This work was supported by Guangxi Key Research and Development Program (No. AB22035056).

Conflict of interests

The authors report no conflict of interest.

Data statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Multimedia Appendix 1:

A sample of chief complain and medical histories.

Multimedia Appendix 2:

Merged concepts and question templates.

Multimedia Appendix 3:

Overview of comparison between LLMs outputs and expert's annotation.

Multimedia Appendix 4:

Overview of comparison between LLMs outputs and expert's annotation when omitting corpus extraction step.

Multimedia Appendix 5

LLMs Q&A scales precision.

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. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2021;13(6):e1549.
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. *Journal of healthcare engineering*. 2018;2018.
3. Varshini KS, Uthra RA, editors. *An Approach to Extract Meaningful Data from Unstructured Clinical Notes*. Inventive Systems and Control: Proceedings of ICISC 2021; 2021: Springer.
4. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform*. 2020 Mar 31;8(3):e17984. PMID: 32229465. doi: 10.2196/17984.
5. Wu PH, Yu A, Tsai CW, Koh JL, Kuo CC, Chen ALP. Keyword extraction and structuralization of medical reports. *Health Inf Sci Syst*. 2020 Dec;8(1):18. PMID: 32269770. doi: 10.1007/s13755-020-00108-6.
6. Vassar M, Matthew H. The retrospective chart review: important methodological considerations. *Journal of educational evaluation for health professions*. 2013;10.
7. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. *The American journal of managed care*. 2002;8(9):787-93.
8. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. *Evaluation the health professions*. 2009;32(3):281-98.
9. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*. 2022 2022 □ 12 □ 22 □;1(12):e0000168. doi: 10.1371/journal.pdig.0000168.
10. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. *Journal of biomedical informatics*. 2009;42(5):923-36.

11. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*. 2020;21(1):5485-551.
12. Devlin J, Chang M-W, Lee K, Toutanova K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 October 01, 2018:[arXiv:1810.04805 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D>.
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*. 2020 Feb 15;36(4):1234-40. PMID: 31501885. doi: 10.1093/bioinformatics/btz682.
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*. 2019 Sep 12;7(3):e14830. PMID: 31516126. doi: 10.2196/14830.
15. Wang S, Yilahun H, Hamdulla A, editors. Biomedical Named Entity Recognition Based on MCBERT. 2022 International Conference on Asian Language Processing (IALP); 2022: IEEE.
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. *Journal of medical Internet research*. 2021;23(1):e19689.
17. Zhang Z, Zhu L, Yu P. Multi-level representation learning for Chinese medical entity recognition: Model development and validation. *JMIR Medical Informatics*. 2020;8(5):e17637.
18. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*. 2018;77:34-49.
19. Polnaszek B, Gilmore-Bykovskiy A, Hovanes M, Roiland R, Ferguson P, Brown R, et al. Overcoming the challenges of unstructured data in multi-site, electronic medical record-based

abstraction. Medical care. 2016;54(10):e65.

20. Tarik Altuncu M, Sorin E, Symons JD, Mayer E, Yaliraki SN, Toni F, et al. Extracting information from free text through unsupervised graph-based clustering: an application to patient incident records 2019 August 01, 2019:[arXiv:1909.00183 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2019arXiv190900183T>.
21. Choi HS, Song JY, Shin KH, Chang JH, Jang B-S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiation Oncology Journal. 2023;41(3):209.
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 Network Open. 2023;6(10):e2336997-e.
23. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI cancer spectrum. 2023;7(2):pkad015.
24. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of Medical Systems. 2023;47(1):33.
25. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. European Journal of Cardiovascular Nursing. 2023:zvad087.
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. Journal of Medical Internet Research. 2023;25:e51580.
27. Thapa S, Adhikari S. ChatGPT, Bard, and Large Language Models for Biomedical Research:

Opportunities and Pitfalls. *Annals of Biomedical Engineering*. 2023:1-5.

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. *Journal of Medical Internet Research*. 2023;25:e49240.
29. Zhang L, Liu X, Li Z, Pan X, Dong P, Fan R, et al. Dissecting the Runtime Performance of the Training, Fine-tuning, and Inference of Large Language Models 2023 November 01, 2023: [arXiv:2311.03687 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv231103687Z>.
30. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A Survey on Large Language Model based Autonomous Agents 2023 August 01, 2023: [arXiv:2308.11432 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230811432W>.
31. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*. 2022;35:24824-37.
32. Lei D, Li Y, Hu M, Wang M, Yun V, Ching E, et al. Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations 2023 October 01, 2023: [arXiv:2310.03951 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv231003951L>.
33. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models 2022 March 01, 2022: [arXiv:2203.11171 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2022arXiv220311171W>.
34. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen Technical Report 2023 September 01, 2023: [arXiv:2309.16609 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230916609B>.
35. Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C, et al. Baichuan 2: Open Large-scale Language Models 2023 September 01, 2023: [arXiv:2309.10305 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230910305Y>.

36. ACOG Committee Opinion No. 743: Low-Dose Aspirin Use During Pregnancy. *Obstet Gynecol.* 2018 Jul;132(1):e44-e52. PMID: 29939940. doi: 10.1097/aog.0000000000002708.
37. National Institute for Health and Care Excellence: Guidelines. Hypertension in pregnancy: diagnosis and management 2019 Available at: www.nice.org.uk/guidance/ng133 London: National Institute for Health and Care Excellence (NICE) Copyright © NICE 2019.; 2020.
38. OpenAI. Openai api. 2023; Available from: <https://openai.com/blog/openai-api>.
39. Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena 2023 June 01, 2023:[arXiv:2306.05685 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230605685Z>.
40. International Health Terminology Standards Development Organisation. SNOMED CT. [Aug 12/2009]; Available from: <http://www.ihtsdo.org/snomed-ct/>.
41. Clowse ME, Jamison M, Myers E, James AH. A national study of the complications of lupus in pregnancy. *American journal of obstetrics gynecology.* 2008;199(2):127. e1-. e6.
42. Rees F, Doherty M, Grainge MJ, Lanyon P, Zhang W. The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. *Rheumatology.* 2017;56(11):1945-61.
43. Hwang J-J, Shin S-H, Kim Y-J, Oh Y-M, Lee S-D, Kim YH, et al. Epidemiology of antiphospholipid syndrome in Korea: a nationwide population-based study. *Journal of Korean Medical Science.* 2020;35(5).
44. Xu X, Rao Y, Wang L, Liu S, Guo JJ, Sharma M, et al. Smoking in pregnancy: a cross-sectional study in China. *Tobacco Induced Diseases.* 2017;15:1-8.
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. *Journal of the American Heart Association.* 2017;6(3):e004924.
46. Leon N, Balakrishna Y, Hohlfeld A, Odendaal WA, Schmidt B-M, Zweigenthal V, et al.

Routine Health Information System (RHIS) improvements for strengthened health system management. Cochrane Database of Systematic Reviews. 2020 (8).

47. Miñarro-Giménez JA, Cornet R, Jaulent M-C, Dewenter H, Thun S, Gøeg KR, et al.

Quantitative analysis of manual annotation of clinical text samples. International journal of medical informatics. 2019;123:37-48.

48. Rawte V, Chakraborty S, Pathak A, Sarkar A, Towhidul Islam Tonmoy SM, Chadha A, et al.

The Troubling Emergence of Hallucination in Large Language Models -- An Extensive Definition, Quantification, and Prescriptive Remediations2023 October 01, 2023:[arXiv:2310.04988 p.].

Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv231004988R>.

49. Jones E, Palangi H, Simões C, Chandrasekaran V, Mukherjee S, Mitra A, et al. Teaching

Language Models to Hallucinate Less with Synthetic Tasks2023 October 01, 2023:

[arXiv:2310.06827 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv231006827J>.

Multimedia Appendix 1

A sample of chief complain and medical histories.

Translated version.

This material contains a synthesized example of a patient's admission record, translated from the original Chinese text, and is entirely fictionalized with no real patient data or identifiable private information included.

Chief Complaint:

Amenorrhea of 36 weeks and 5 days, vaginal fluid discharge for over 1 hour.

Present Illness History:

The patient complains of historically irregular menstruation, 6-7/37-45 days, with the last menstrual period on December 31, 2019. On March 18, 2020, a B-ultrasound of the uterus and adnexa in our hospital showed: intrauterine early pregnancy, visible fetal heartbeat, and a small cyst in the right adnexal area measuring about 15×10mm. According to B-ultrasound, the revised expected date of delivery is October 24, 2020. On March 27, 2020, a health check-up card was established at our hospital, with an initial diagnosis blood pressure of 98/70mmHg. The patient experienced significant early pregnancy symptoms like nausea and fatigue, which disappeared on their own after about 3 months of pregnancy. During the health check-up, tests for routine blood and urine, liver and kidney functions, electrolytes, coagulation function, eugenics eight items, hepatitis B two and a half pairs, immune four items, glucose-6-phosphate dehydrogenase, vaginal discharge routine, Ureaplasma urealyticum + gonococcus + Chlamydia trachomatis DNA determination showed no significant abnormalities. Free thyroid function three items: TSH 5.16 uIU/ml; Free T4 11.05 pmol/L; Free T3 5.48 pmol/L. The endocrinology outpatient clinic of our hospital diagnosed "pregnancy with hypothyroidism" and prescribed oral levothyroxine 1.25 tablets to treat hypothyroidism. Thyroid function was monitored during pregnancy and was acceptable. Both partners screened negative for thalassemia. On April 17, 2020, fetal NTB ultrasound showed: intrauterine pregnancy, single live fetus, early pregnancy, equivalent to 12 weeks and 6 days of gestation, CRL 65mm, NT 0.8mm. Mid-term Down syndrome screening all indicated low risk. From about 4 months of pregnancy, the patient began to feel fetal movements, which continue to the present. On June 30, 2020, a fetal four-dimensional color ultrasound at our hospital indicated: BPD 47mm, HC 183mm, a single live fetus in the uterus, mid-pregnancy, equivalent to 22 weeks of gestation; no significant complex malformations were observed in the fetal heart; fetal umbilical blood flow ratio showed no significant abnormalities. On July 20, 2020, a fetal B-ultrasound at our hospital indicated: BPD 56mm, HC 218mm, a single live fetus in the uterus, mid-pregnancy, breech position, according to various fetal biometric measurements, equivalent to 24 weeks and 5 days of gestation, biparietal diameter equivalent to 23 weeks and 1 day of gestation; no significant complex malformations were observed in the fetal heart; fetal umbilical blood flow ratio showed no significant abnormalities; placental function level 0; the fetal umbilical cord was wrapped around the neck twice. Because the biparietal diameter was less than 2 weeks of gestational age, umbilical blood puncture was performed at 26 weeks of pregnancy for chromosomal karyotype analysis, which showed no abnormalities. OGTT test was performed: 4.45-10.21-9.45mmol/l. Diagnosed with "gestational diabetes", dietary control was administered under the guidance of our hospital's nutrition department, and blood sugar control was acceptable. On August 20, 2020, a fetal system B-ultrasound at our hospital indicated: BPD 67mm, HC 275mm, AC

258mm, HL 51mm, FL 57mm, the placenta is located on the posterior wall of the uterus, the maximum anteroposterior diameter of the amniotic fluid is 47mm. A single live fetus in the uterus, late pregnancy, breech position, according to various fetal biometric measurements, equivalent to 29 weeks and 3 days of gestation; no significant complex malformations were observed in the fetal heart; fetal umbilical blood flow ratio showed no significant abnormalities; placental function level 0; the fetal umbilical cord was wrapped around the neck once. GBS test was performed in late pregnancy, indicating negative. Around 20:40 on October 1, 2020, a small amount of vaginal fluid was discharged, occasional abdominal distension was noted, no abdominal pain, no vaginal bleeding, self-perceived normal fetal movements, and the patient is now visiting our hospital. There have been no recent discomforts such as chills, fever, dry cough, sore throat, nasal congestion, runny nose, diarrhea, muscle aches, etc. The emergency department's preliminary diagnosis is "1. Premature rupture of membranes 2. 36 weeks + 5 days of gestation (G1P0) 3. Vertex presentation 4. Gestational diabetes 5. Pregnancy with hypothyroidism 6. Umbilical cord entanglement" and the patient was admitted to the hospital. During pregnancy, the patient denies exposure to toxins and radiation, denies chills, fever, dizziness, blurred vision, denies abdominal distension, abdominal pain, denies chest tightness, palpitations, and other discomforts, is in good spirits, has a good appetite, sleeps well, has normal bowel movements, and gained 13kg during pregnancy.

Past History:

Previously healthy, denies a history of hypertension, diabetes, blood diseases, and heart disease. Denies history of hepatitis and other infectious diseases like tuberculosis, denies significant trauma history, denies history of blood transfusion, denies food and drug allergies, vaccination history is unclear, and review of other systems revealed no abnormalities.

Personal History:

Born and resides in her native place, denies long-term residence history in other places, denies bad habits like smoking and drinking.

Menstrual and Obstetric History:

Menarche at 13 years old, menstrual period of 6-7 days, menstrual cycle of 37-45 days, last menstrual period: December 31, 2019. The menstrual volume is medium, with dark red color, and no dysmenorrhea. Has a steady sexual partner, G1P0.

Family History:

Parents are healthy, denies family history of genetic diseases and infectious diseases.

Original Chinese version.

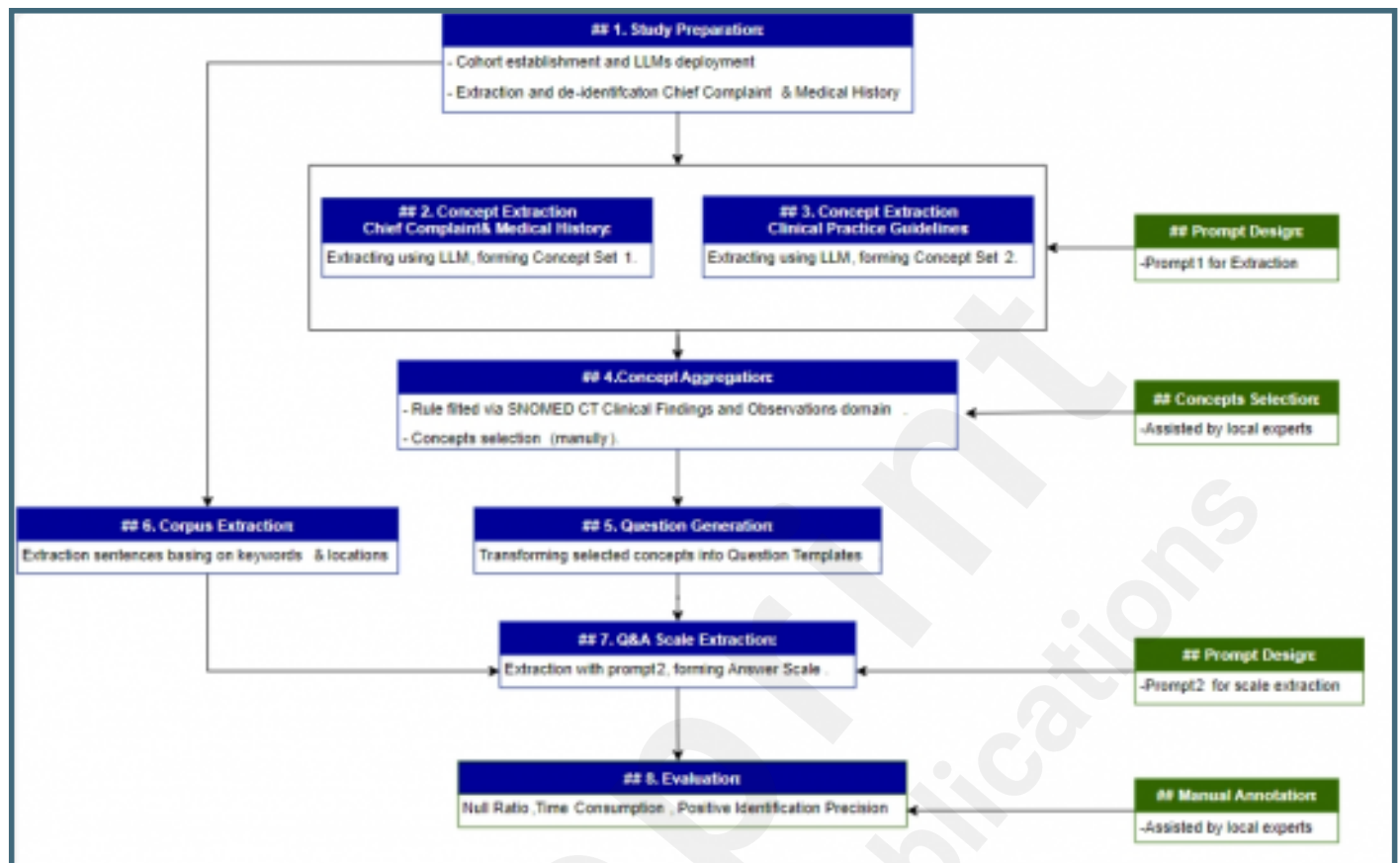
孕周 36 周 5 天 孕产史 1 产
 孕周 6-7/37-45 孕产史 2019 年 12 月 31 日 2020 年 3 月 18 日 孕产史 B 孕产史 15×10mm 孕产史 B 孕产史 2020 年 10 月 24 日 2020 年 3 月 27 日 孕产史 98/70mmHg 孕产史 3 孕产史 -6- 孕产史 + 孕产史 + 孕产史 DNA 孕产史 TSH 5.16 uIU/ml 孕产史 T4 11.05 pmol/L 孕产史 T3 5.48 pmol/L 孕产史 "孕产史" 孕产史 1.25 孕产史 2020 年 4 月 17 日 孕产史 NTB 孕产史 12 孕产史 +6 孕产史 CRL 65mm 孕产史 NT 0.8mm 孕产史 4 孕产史 2020 年 6 月 30 日 孕产史 BPD 47mm 孕产史 HC 183mm 孕产史 22 孕产史 2020 年 7 月 20 日 孕产史 B 孕产史 BPD 56mm 孕产史 HC 218mm 孕产史 24 周 5 天 孕产史

23 1 0 2 26 OGTT 4.45-10.21-9.45mmol/l " 2020 8 20 B BPD 67mm HC 275mm AC 258mm HL 51mm FL 57mm 47mm 29 3 0 GSB 2020 10 1 20:40 "1. 2.36 +5 (1 0) 3. 4. 5. 6." 13 6-7 37-45 2019 12 31 1 0

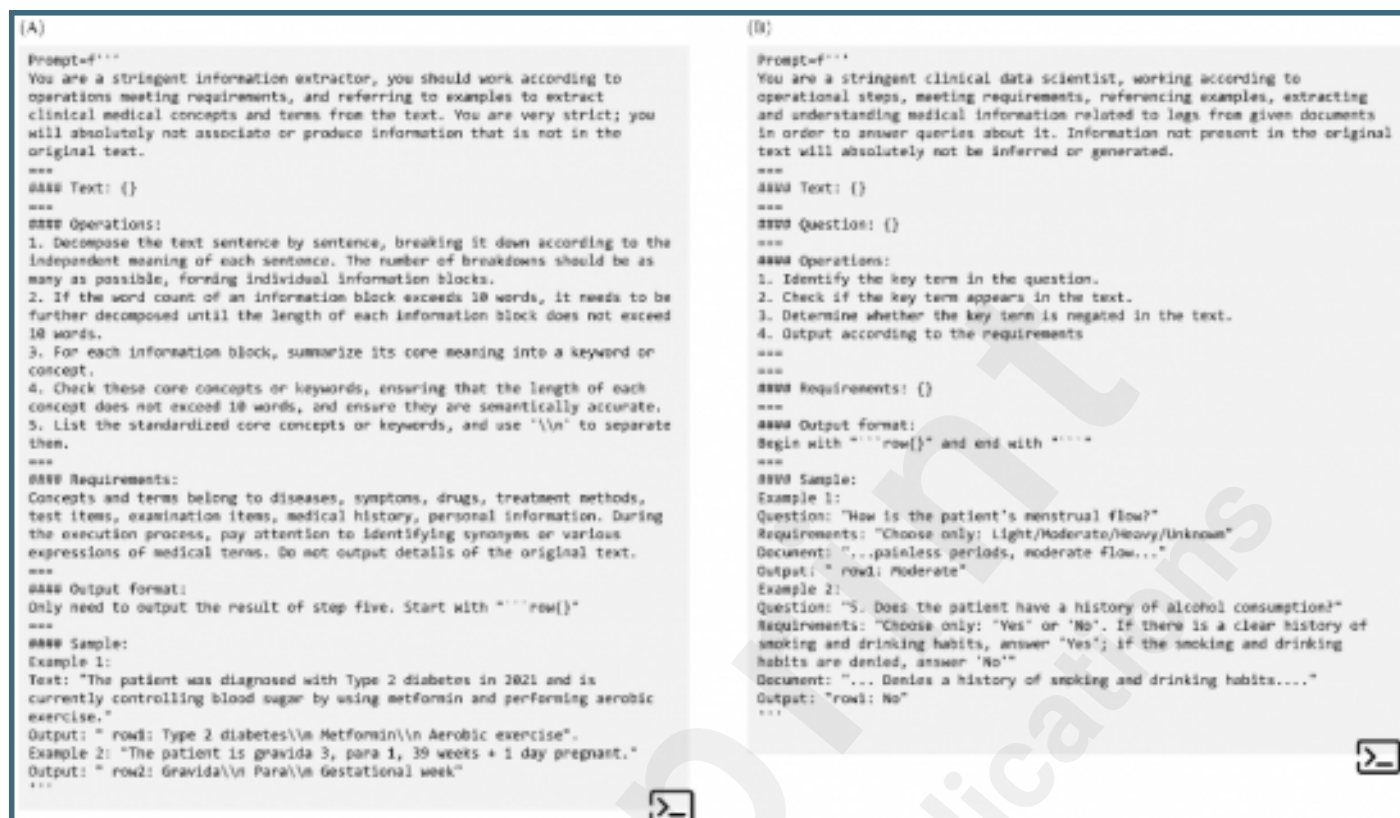
Supplementary Files

Figures

Flow chart of Extraction approach.

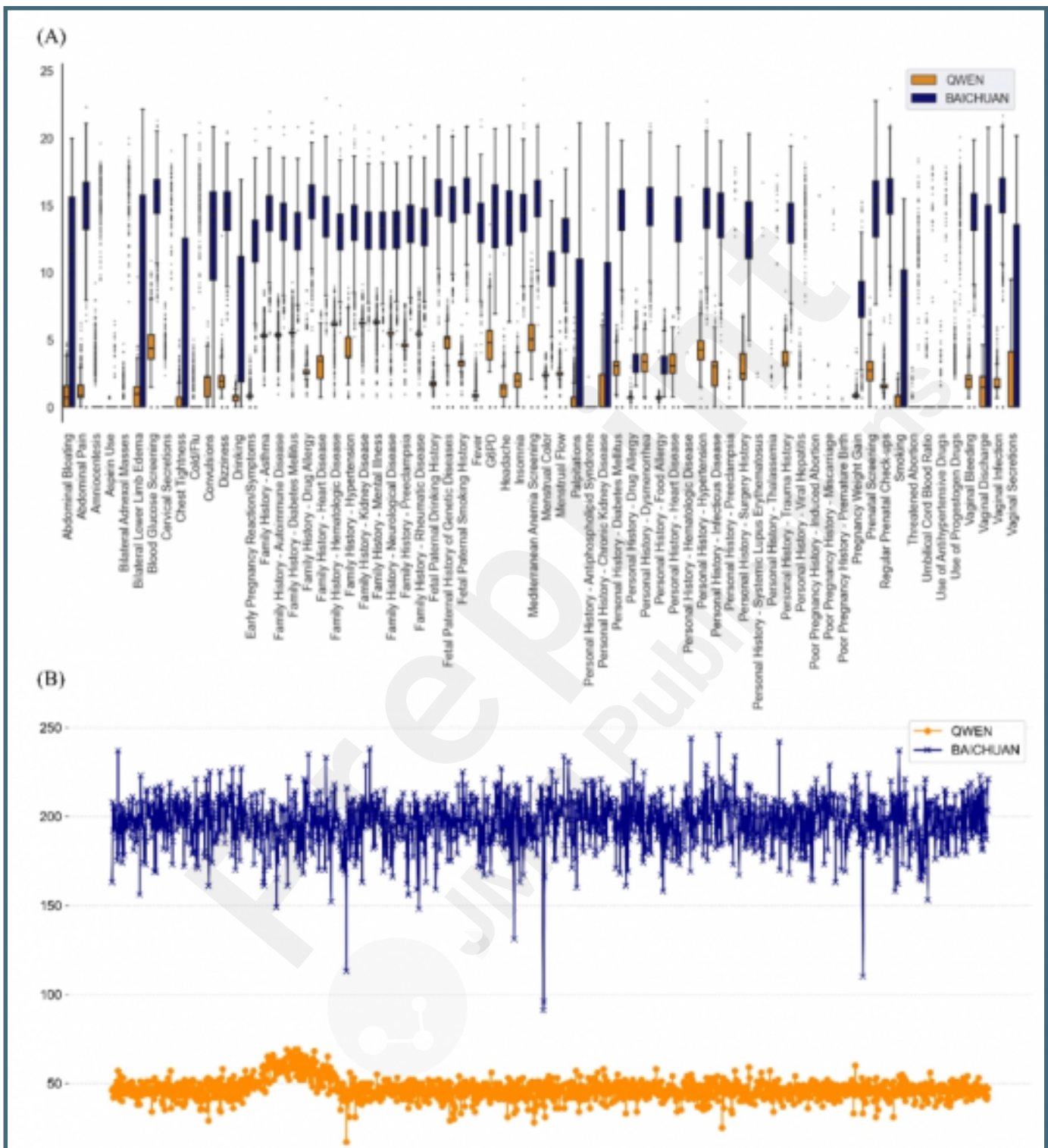


Prompt templates of information extraction and Q&A scales.



[illegible]

Time Consumption of Q&A scales for QWEN and BAICHUAN, measured in seconds.



Approach performance when omitting corpus extraction step.



Multimedia Appendixes

A sample of chief complain and medical histories.

URL: <http://asset.jmir.pub/assets/3e0c713d51bffb1b9eb409887e4d9449.docx>

Merged concepts and question templates.

URL: <http://asset.jmir.pub/assets/b0756f023ca30574c58285dfed5fbb92.docx>

Supplementary Figure 1. Overview of comparison between LLMs outputs and expert's annotation.

URL: <http://asset.jmir.pub/assets/abfcabd6f5d0642264237bb5692ad408.pdf>

Supplementary Figure 2. Overview of comparison between LLMs outputs and expert's annotation when omitting corpus extraction step.

URL: <http://asset.jmir.pub/assets/66d7ee0dfaf76a4319c24e1f0006bb07.pdf>

LLMs Q&A scales precision.

URL: <http://asset.jmir.pub/assets/a40c517a507f8bec5df713eb65f67e11.docx>