# Comparative Analysis of Large Language Models in Traditional Chinese Medicine

**Authors:**

Xiao-Zhou Lu[a#], M.M., Hang-Tong Hu[b#], M.D., Ting Xiang[a#], M.D., Jin-Zhen Wu[a], M.M., Peng Liu[a], M.M., Wei-Ming Ji[a], M.M., Guan-Heng He[c], M.D., Ze-Xiong Chen[a*], , M.M., Wei Wang[b*], M.D., Ph.D., Bao-Guo Sun[a*], M.D.

[a] Department of Traditional Chinese Medicine, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[b] Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, Ultrasomics Artificial Intelligence X-Lab, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou, China.

[c] Department of Acupuncture, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

[#] Xiao-Zhou Lu, Hang-Tong Hu and Ting Xiang contributed equally to the study.

***Correspondence Author:**

Bao-Guo Sun, MD,

Department of Traditional Chinese Medicine, The First Affiliated Hospital, Sun Yat-Sen University, 58 Zhongshan Road 2, Guangzhou, 510080, People's Republic of China.

Phone and Fax: 86-20-87765183

E-mail: sunbaog@mail.sysu.edu.cn

Wei Wang, MD, PhD,

Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, Ultrasomics Artificial Intelligence X-Lab, The First Affiliated Hospital, Sun Yat-Sen University, 58 Zhongshan Road 2, Guangzhou, 510080, People's Republic of China.

Phone and Fax: 86-20-87765183

E-mail: wangw73@mail.sysu.edu.cn

Ze-Xiong Chen, MM,

Department of Traditional Chinese Medicine, The First Affiliated Hospital, Sun Yat-Sen University, 58 Zhongshan Road 2, Guangzhou, 510080, People's Republic of China.

Phone and Fax: 86-20-87765183

E-mail: 13711191912@163.com

## Research in context

### Evidence before this study

In recent years, Traditional Chinese Medicine (TCM), with its history spanning over two millennia, has increasingly gained recognition on the international stage. However, its unique logical framework and complex diagnostic and therapeutic processes pose significant challenges for clinicians and students. These factors have limited its broader application and effectiveness in patient management. With the rapid advancement of artificial intelligence, particularly the emergence of large language models (LLMs), there has been growing interest in their application in the medical field. From medical examinations to clinical practice, LLMs have demonstrated their potential in assisting diagnosis, knowledge dissemination, and learning assessment, offering new opportunities for the advancement of TCM. However, TCM's distinctive linguistic and contextual features remain a challenge for LLMs. In this study, we aim to comprehensively evaluate the performance of chatbot-based LLMs in mastering TCM knowledge and applying it in clinical practice. Web of Science searches for those terms related to this approach performed up to March, 2024, with no language restrictions, did not reveal any publications in this area.

### Added value of this study

We evaluated eight LLMs, including those developed in both Western countries and China, on their foundational knowledge of TCM using a combination of simulated and actual questions from the TCM licensing examination. The best-performing LLMs were then selected for further evaluation in syndrome diagnosis and herbal formula generation, using 100 real clinical cases. Their performance in clinical practice was compared against that of experienced physicians. The results demonstrated that Baidu's ERNIE series and ChatGPT 4o have a strong foundation in TCM and excel in examinations, achieving diagnostic performance comparable to that of professional physicians. However, they exhibited significant limitations in generating accurate herbal formula recommendations.

**Implications of all the available evidence**

For the first time, we conducted a comprehensive evaluation of LLMs in mastering TCM knowledge, as well as their performance in clinical diagnosis and herbal formula generation. While certain LLMs performed well in TCM examinations, most showed a relatively limited understanding of classical TCM texts. Notably, the ERNIE series and ChatGPT 4o demonstrated strong analytical capabilities in diagnosis, yet their accuracy in herbal formula recommendations remained insufficient. Despite these limitations, LLMs show significant promise for applications in TCM, and future research should focus on targeted training to address these areas of weakness.

## Abstract

**Background:** Traditional Chinese Medicine (TCM), with its unique framework and linguistic characteristics, presents challenges for large language models (LLMs). Identifying LLMs' strengths and limitations is crucial to exploring their potential in the TCM domain. Herein, we evaluated the performance of multiple LLMs in TCM, including their foundational knowledge, diagnostic capabilities, and herbal prescription generation.

**Methods:** This comparative research was conducted in two phases: (1) assessing LLMs' mastery of TCM foundational knowledge with two question banks (1911 questions in Question Bank A and 100 questions in Question Bank B) that were accessed from a paywalled online platform. (May–June 2024); and (2) evaluating LLMs' syndrome differentiation and prescription generation using 100 real-world clinical cases collected from the First Affiliated Hospital of Sun Yat-sen University (July–August 2024). Eight LLMs (ChatGPT 3.5, ChatGPT 4o, Gemini, Gemini Advanced, ERNIE 3.5, ERNIE 4, GLM 3, and GLM 4) were evaluated. High-performing models advanced to the second phase, and their outputs were compared with diagnoses and prescriptions from TCM practitioners. LLMs' accuracy was assessed by percentage scores and consistency by Intraclass Correlation Coefficient (ICC). TCM syndrome diagnoses, decomposed into disease location and nature elements, were evaluated by the Dice Similarity Coefficient (DSC), and herbal prescription quality was assessed by expert review.

**Findings:** In the examination, ERNIE 4 and ERNIE 3.5 scored the highest (78.6%), followed by ChatGPT 4o (76.6%). ERNIE 4 achieved the highest ICC (0.92), with ERNIE 3.5 (0.89) and ChatGPT 4o (0.78) also demonstrating strong repeatability. In clinical cases, ChatGPT 4o outperformed ERNIE 3.5 in syndrome differentiation for both disease location (DSC: 0.74 vs. 0.63, $p$<.05) and disease nature

(DSC: 0.73 vs. 0.64, $p$<.05), and surpassed ERNIE 4 in disease nature differentiation (DSC: 0.73 vs. 0.64, $p$<.05). ChatGPT 4o also achieved the highest prescription score (119), though only 45% of its prescriptions fully matched the reference.

**Interpretation:** LLMs excelled in TCM knowledge and diagnosis but need improvement in generating effective herbal prescriptions.

**Funding:** National Nature Science Foundation of China (NO: 82205248 and NO: 82272076) and Guangdong Natural Science Foundation (NO：2022A1515012133).

**Keywords:** Traditional Chinese Medicine; large language models chatbots; foundational knowledge; syndrome diagnosis; prescription generation

## Introduction

Traditional Chinese Medicine (TCM), a time-honored medical system characterized by its unique philosophical foundation and therapeutic approaches distinct from modern biomedicine, has gained growing international recognition in recent years(1). Even though advanced diagnostic instruments and technologies have emerged nowadays, TCM has preserved its distinctive characteristic of syndrome differentiation through the four diagnostic methods: inspection, auscultation and olfaction, inquiry, and palpation. This process demands the integration and interpretation of various forms of information and reasoning based on the patient's symptoms and signs, making it difficult for less experienced practitioners to master. Furthermore, TCM has continuously evolved over its long history, giving rise to numerous schools of thought. As a result, different practitioners may adopt varying treatment strategies for the same illness. The complexity and diversity of TCM diagnostic and treatment procedures present challenges in patient management, as well as in the popularization and education of TCM.

Large Language Models (LLMs), a revolutionary advancement of Natural Language Processing (NLP), emerges as a potential solution for TCM's challenges. LLMs can generate human-like responses and interact seamlessly with users(2). With the rapid development of Artificial Intelligence (AI) techniques and increased computational resources, LLMs have permeated various sectors, including healthcare(3). Over the past two years, LLMs like ChatGPT have garnered significant attention by passing the threshold of the United States Medicine Licensing Exam (USMLE) (4). Notably, GPT-4 surpassed the average human score on specialized medical board examinations, such as neurology(5), highlighting the potential for AI-assisted learning in medical education. There is also increasing research focused on applying LLMs in clinical practice. These studies demonstrated that LLMs could automatically and accurately generate structured reports from original radiology reports (6, 7), diagnose complex clinical cases(8), and provide empathetic, high-quality responses to patient inquiries(9) by leveraging their ability to analyze and extract valuable information from textual content. The advantage of rapid assimilation and summarization of information allows LLMs to explore the rich resources of TCM literature and clinical data(10), offering students and young doctors a powerful tool for enhancing their learning and clinical skills. Additionally, the emergence of multimodal LLMs, which can integrate textual and non-textual data like images and audio(11), can further assist in the real-time analysis of patient cases and help standardize and optimize the diagnostic process in TCM.

In recent years, significant progress has been made in developing language models for the TCM domain(12-14), with efforts aimed at bridging the gap between ancient wisdom and modern technology.

For instance, Hua et al(15). constructed datasets incorporating extensive TCM theoretical and clinical knowledge and developed a large language model tailored for TCM, which achieved superior results in various tasks such as clinical Q&A and herbal prescription recommendation. Tan et al.(16) introduced MedChatZH, a dialogue model optimized for TCM consultations, which performed effectively on real-world medical dialogue datasets. However, unlike modern biomedicine, which relies on standardized terminologies and clear structures rooted in anatomy and biology, TCM is based on the holistic concept and ancient Chinese philosophical ideas mainly derived from Taoism, such as *yin*, *yang*, and *qi.* The theory of the five elements, in particular, is often abstract and difficult to quantify. Besides, the language of TCM is metaphorical and context-dependent, with terms like “liver fire” or “liver wind” representing imbalances related to complex pathologies and symptoms rather than their literal meanings. The divergence in medical frameworks and linguistic nuances presents challenges for LLMs, primarily trained on structured, standardized data from modern medicine and modern language corpora, to fully interpret and apply TCM knowledge(17). Though LLMs have shown promise in theoretical and controlled experimental settings, their effectiveness and safety in real-world TCM clinical environments remain largely unverified.

This study aims to first assess LLMs’ understanding of foundational TCM knowledge through the use of TCM Practitioner Licensing Examination questions, followed by an evaluation of their diagnostic and prescription generation capabilities. In a novel approach, we utilize real-world clinical cases as input for the LLMs and compare their outputs with the diagnoses and prescriptions provided by experienced TCM practitioners.

## Materials and Methods

Although TCM operates within a complex reasoning system, it is still grounded in fundamental theoretical knowledge. Thus, the first part of our study was designed to assess the performance of LLMs in foundational knowledge through their responses to questions from the TCM licensing examination. The second part, focused on clinical case analysis, evaluated the models' inductive reasoning abilities, particularly their proficiency in TCM-specific reasoning, as demonstrated by their syndrome differentiation and prescription generation performance. The research workflow is shown in Figure 1. The retrospective analysis obtained ethical approval (No.2021464) and waived the informed consent requirement.

**Large Language Models**

Considering that language and cultural biases may affect LLMs' performance in the context of TCM(18), we simultaneously selected models developed by both Western countries and Chinese companies. Eight LLMs were evaluated. They were: ChatGPT 3.5 and ChatGPT 4o developed by OpenAI (https://chatgpt.com), Gemini and Gemini advanced by Google (https://gemini.google.com/app), Ernie Bot 3.5 and Ernie Bot 4 by Baidu (https://yiyan.baidu.com/), ChatGLM-3 and ChatGLM-4 by ZHIPU AI ( https://chatglm.cn/main/alltoolsdetail ). Characteristics of the models are listed in Supplement 1. These models have been trained on a large corpus of text data, which enabled them to interact with humans in a conversational way.

**Preparation of the datasets**

1. Multiple-Choice Question Bank

The National Medical Licensing Examination for Traditional Chinese Medicine is the qualification entrance exam for TCM practitioners in China. It consists of multiple-choice questions covering 18 subjects including Fundamentals of TCM, Diagnostics of TCM, Chinese Materia Medica, Chinese Medical Formulae, Yellow Emperor's Inner Canon, Treatise on Cold Damage, Essential Prescriptions from the Golden Cabinet, Warm Disease Studies, Internal Medicine of TCM, External Medicine of TCM, Gynecology of TCM, Pediatrics of TCM, Acupuncture and Moxibustion, Fundamentals of Diagnostics, Internal Medicine, Infectious Diseases, Medical Ethics, Laws and Regulations. To comprehensively analyze the LLMs' performance across various subjects, we created a question bank covering all the above subjects. We first subscribed to a question bank from an educational company (https://www.qingsongxueba.com) of 6370 multiple-choice questions that resemble the National Medical Licensing Examination for TCM Practitioners. Each subject in the question bank consisted of different amounts of questions following the exam's proportion. We then randomly selected 30% of questions from each subject and created the *TCM Examination Question Bank A* with 1911 questions to test the LLMs' competence. We also randomly selected 100 questions from the real exam in the year 2022, the *TCM Examination Question Bank B,* to test the LLMs' repeatability. All the questions had only one best answer from 5 options: A, B, C, D, and E.

2. Medical cases collection

We accessed the Haitai Electronic Medical Record System of the First Affiliated Hospital of Sun Yat-sen University to collect the medical histories of 100 outpatients from December 2023 to July 2024. To guarantee the reliability and effectiveness of the prescriptions delivered by the doctors, the medical cases inclusion criteria were as follows: (a) the attending doctor was a senior professional with over 15 years of clinical experience ; (b) the medical history was thoroughly recorded, including the main complaints, symptoms, signs, diagnosis (both disease diagnosis and TCM syndrome differentiation), and Chinese medicine prescription; (c) the therapeutic effect was confirmed from the patient's second visit record with notable improvement in symptoms based on clinical observations.

The TCM syndrome, also known as "Zheng", refers to a collection of clinical symptoms and signs reflecting the imbalance of the body's internal organs, Qi, or blood, at a certain stage(19). Syndrome elements (SEs), the fundamental components of TCM syndrome, summarize a disease's cause, nature, and location, serving as the foundation for its treatment strategy(20). By breaking down the TCM syndrome diagnoses into syndrome elements, we can not only better understand the disease's underlying characteristics but also standardize TCM syndromes for assessing the diagnostic capabilities of LLMs. The 100 cases of TCM syndromes diagnosed by the doctors were decomposed into pathological location elements (e.g., spleen, liver) and pathological nature elements (e.g., Qi deficiency, blood stasis).

**Section One: LLMs' Performance in the TCM Examination**

We logged into the official websites of each language model's chatbot. We then presented the following prompt: "You are about to take the licensing examination for TCM practitioners. Your task is to select the correct answer from the following single-choice questions, with options labeled A, B, C, D, and E". After providing this prompt, we inputted ten questions at a time and recorded the models' responses. For the *TCM Examination Question Bank A*, we recorded each model's total score as well as its score in each subject based on correct answers. The scores were converted to percentages to facilitate comparison across subjects and to determine which models would advance to section two. To test the LLMs' repeatability, we used *TCM Examination Question Bank B*, which consisted of 100 questions. Each model was given the same set of questions five times at different intervals, and their answers were collected. The examination section took place from May 8 to June 9, 2024.

**Section Two: LLMs' Diagnosis and Prescription Generation Performance**

The well-performed chatbots in the TCM examination underwent evaluation of their diagnosis and prescription generation abilities. We first provided a prompt to the LLM chatbot: "Now you are a professional TCM doctor, and you must make a syndrome differentiation diagnosis and generate a

Chinese medicine formula containing the herbs and dosages according to the patient's medical information. When diagnosing, you must simultaneously output the syndrome elements of pathological location and nature. " After getting a positive response from the LLM, we inputted the patient's gender, age, and medical history, such as symptoms and the description of tongue and pulse. We concealed the patient's name and medical record number to prevent personal information leakage.

We compared the model's syndrome differentiation results with those of the doctors using the Dice Similarity Coefficient (DSC), a statistical measure to gauge the similarity between two data sets(21). To assess the treatment effectiveness of LLMs, we invited a TCM senior physician (working experience of 20 years) to review the herbal formulas prescribed by the models, using the attending doctor's prescriptions as the reference standard. The evaluator rated the conformity of the prescriptions from the model and the doctor in terms of treating strategy and composition similarity on a three-point scale (where 0 = do not match, 1 = partially matches, and 2 = almost fully matches). The attending doctors of the 100 cases were excluded when we selected the evaluator. A blind method was adopted to prevent bias, ensuring the evaluator was unaware of which models had generated the formulas. Supplement 2 shows an example of the scoring sheet displayed for the evaluator. In our study, the prompts and the outputs were in Chinese only.

**Statistical Analysis**

Descriptive statistics were summarized as the mean ± standard deviation (SD) or median and interquartile range. The Intraclass Correlation Coefficient (ICC) assessed the consistency of LLMs' responses. As previously mentioned, the Dice Similarity Coefficient measured the overlap between two text documents, ranging from 0 (no similarity) to 1 (identical). The t-test was employed to compare continuous variables, while the Wilcoxon Signed-Rank Test was used to compare ordinal data. Results with two-sided *P*-values of less than 0.05 indicated a statistically significant difference. All the statistical analyses were conducted using Python 3.12.5 (https://www.python.org). Specifically, the t-test and the Wilcoxon Signed-Rank Test were performed using the "scipy.stats" package, the ICC was calculated using the "pingouin" package, and the DSC was computed with the "sklearn.metrics" package.

## Results

### Section One: LLMs' Performance in the TCM Examination

The *TCM Examination Question Bank A* was to test LLMs' proficiency in TCM knowledge. Table 1 shows the details of *Question Bank A* and LLMs' scores of each subject. We ranked the LLMs according to their total scores (the last row in Table 1) in descending order: ERNIE Bot 4 (78.6%), ERNIE Bot 3.5 (78.6%), ChatGPT 4o (76.6%), GLM 4 (70.0%), Gemini Advanced (69.2%), GLM 3 (65.5%), Gemini (42.4%), ChatGPT 3.5 (39.0%). Although ERNIE Bot 4 and ERNIE Bot 3.5 got the same score, we regarded ERNIE Bot 4 as the benchmark and compared it with the others. Among them, there was no statistical difference between ERNIE Bot 3.5 and ERNIE Bot 4, and ChatGPT 4o and ERNIE Bot 4, while GLM 4, Gemini Advanced, GLM 3, Gemini, and ChatGPT 3.5 were significantly inferior to ERNIE Bot 4 (Figure 2).

We then proceeded to observe each model's performance across 18 subjects. Four models received the lowest score on the subject "Shanghan Lun (Treatise on Cold Damage)": ChatGPT 3.5 (23.8%), Gemini (23.8%), Gemini Advanced (42.9%), and GLM 3 (52.4%). Meanwhile, ERNIE Bot 4 and GLM 4 performed poorly in "Acupuncture and Moxibustion", scoring only 60.6% and 55.9%, respectively. ChatGPT 4o had the lowest score of 52.5% in "Wenbing Xue (Warm Disease Studies)", while ERNIE Bot 3 scored 54.2% in "Huangdi Neijing (Yellow Emperor's Inner Canon)", which was below average for this subject. Among 18 subjects, “Shanghan Lun (Treatise on Cold Damage)” got the lowest average score of 50.5%, while “Internal Medicine” got the highest of 80.2% for all the models. A heatmap displays the models’ accuracy across various subjects (Figure 3). In the heatmap, the color scale ranges from cool to warm colors, where cooler colors (blue) indicate lower scores and warmer colors (red) indicate higher scores. For instance, ChatGPT 4o performs better in “Chinese Materia Medica” and “Internal Medicine” as indicated by the deep red color, while models like ChatGPT 3.5 and Gemini show lower scores in these subjects, represented by cooler colors.

**Table 1. Score Sheet of LLMs in TCM Question Bank A**

| Module | Subject | Question, No. | Average score per subject (%) |
|---|---|---|---|
| **Traditional Chinese Medicine Basics** | Fundamentals of Traditional Chinese Medicine | 112 | 76.9 (68.6) |
| | Diagnostics of Traditional Chinese Medicine | 127 | 80.3 (63.2) |
| | Chinese Materia Medica | 228 | 159.8 (70.0) |
| | Chinese Medical Formulae | 118 | 71.4 (60.5) |
| **Classics of Traditional Chinese Medicine** | Yellow Emperor's Inner Canon | 24 | 13.6 (56.8) |
| | Treatise on Cold Damage | 21 | 10.5 (50.5) |
| | Essential Prescriptions from the Golden Cabinet | 21 | 13.5 (64.3) |
| | Warm Disease Studies | 19 | 11.5 (60.5) |
| **Clinical Practice of Traditional Chinese Medicine** | Internal Medicine of Traditional Chinese Medicine | 224 | 145.1 (64.8) |
| | External Medicine of Traditional Chinese Medicine | 106 | 65.9 (62.2) |
| | Gynecology of Traditional Chinese Medicine | 204 | 129.0 (63.2) |
| | Pediatrics of Traditional Chinese Medicine | 177 | 109.5 (61.9) |
| | Acupuncture and Moxibustion | 170 | 91.4 (53.8) |
| **Western Medicine Integration** | Fundamentals of Diagnostics | 105 | 74.3 (70.7) |
| | Internal Medicine | 116 | 93.0 (80.2) |
| | Infectious Diseases | 87 | 58.6 (68.8) |
| **Medical Humanities** | Medical Ethics | 14 | 9.4 (67.0) |
| | Laws and Regulations | 38 | 28.4 (74.7) |
| **Total score (%)** | | 1911 | |

**Table 1. Score Sheet of LLMs in TCM Question Bank A (Continued)**

| Correct answers, No. (%) | | | | | | | |
|---|---|---|---|---|---|---|---|
| GPT3.5 | GPT 4o | Gemini | Gemini Adv. | ERNIE 3.5 | ERNIE 4 | GLM-3 | GLM-4 |
| 47 (42.0) | 82 (73.2) | 53 (47.3) | 76 (67.9) | 91 (81.2) | 94 (83.9) | 87 (77.7) | 85 (75.9) |
| 46 (36.2) | 91 (71.7) | 54 (42.5) | 87 (68.5) | 101 (79.5) | 97 (76.4) | 77 (60.6) | 89 (70.1) |
| 69 (30.3) | 204 (89.5) | 93 (40.8) | 166 (72.8) | 195 (85.5) | 193 (84.5) | 173 (75.9) | 185 (81.1) |
| 30 (25.4) | 87 (73.7) | 42 (35.6) | 76 (64.4) | 91 (77.1) | 93 (78.8) | 67 (56.8) | 85 (72.0) |
| 12 (50.0) | 14 (58.3) | 11 (45.8) | 13 (54.2) | 13 (54.2) | 16 (66.7) | 16 (66.7) | 14 (58.3) |
| 5 (23.8) | 13 (61.9) | 5 (23.8) | 9 (42.9) | 13 (61.9) | 15 (71.4) | 11 (52.4) | 13 (61.9) |
| 8 (38.1) | 16 (76.2) | 8 (38.1) | 11 (52.4) | 16 (76.2) | 18 (85.7) | 17 (81.0) | 14 (66.7) |
| 10 (52.6) | 10 (52.5) | 8 (42.1) | 14 (73.7) | 14 (73.7) | 12 (63.2) | 12 (63.2) | 12 (63.2) |
| 83 (37.0) | 176 (78.6) | 64 (28.6) | 161 (71.9) | 190 (84.8) | 192 (85.7) | 132 (58.9) | 163 (72.8) |
| 44 (42.0) | 76 (71.7) | 42 (39.6) | 68 (64.2) | 84 (79.2) | 82 (77.4) | 68 (64.2) | 63 (59.4) |
| 85 (41.7) | 144 (70.6) | 85 (41.7) | 125 (61.3) | 160 (78.4) | 157 (77.0) | 137 (67.2) | 139 (68.1) |
| 60 (34.0) | 134 (75.5) | 81 (45.8) | 129 (72.9) | 135 (76.3) | 129 (72.9) | 101 (57.1) | 107 (60.5) |
| 56 (32.9) | 120 (70.6) | 64 (37.6) | 97 (57.1) | 105 (61.8) | 103 (60.6) | 91 (53.5) | 95 (55.9) |
| 44 (41.9) | 89 (84.8) | 57 (54.3) | 80 (76.2) | 83 (79.0) | 84 (80.0) | 86 (81.9) | 71 (67.6) |
| 66 (56.9) | 101 (87.1) | 73 (62.9) | 98 (84.5) | 105 (90.5) | 105 (90.5) | 95 (81.9) | 101 (87.1) |
| 49 (56.3) | 64 (73.6) | 38 (43.7) | 67 (77.0) | 66 (75.9) | 71 (93.1) | 51 (58.6) | 63 (72.4) |
| 7 (50.0) | 10 (71.4) | 8 (57.1) | 11 (78.6) | 10 (71.4) | 10 (71.4) | 10 (71.4) | 9 (64.3) |
| 25 (65.8) | 32 (84.2) | 24 (63.2) | 34 (89.5) | 31 (81.6) | 32 (84.2) | 20 (52.6) | 29 (76.3) |
| 746 (39.0) | 1463 (76.6) | 810 (42.4) | 1322 (69.2) | 1503 (78.6) | 1503 (78.6) | 1251 (65.5) | 1337 (70.0) |

The ICC value of each language model in the *TCM Examination Question Bank B* assessed their repeatability when answering the same set of questions. Among the models, ERNIE Bot 4 demonstrated the highest consistency with an impressive ICC value of 0.92 (95% CI: 0.89–0.94), followed closely by ERNIE Bot 3 with an ICC value of 0.89 (95% CI: 0.85–0.92). ChatGPT 4o came in third with an ICC of 0.78 (95% CI: 0.71–0.83), and Gemini Advanced the fourth with an ICC of 0.75 (95% CI: 0.69–0.81). GLM 4 and GLM 3 achieved ICC values of 0.72 (95% CI: 0.65–0.79) and 0.71 (95% CI: 0.64–0.74), respectively. On the lower end, Gemini exhibited an ICC of 0.62 (95% CI: 0.54–0.70), while ChatGPT 3.5 showed the least consistency with an ICC of 0.53 (95% CI: 0.44–0.62). Results are illustrated in Figure 4.

**Section Two: LLMs' Diagnosis and Prescription Generation Performance**

After comprehensively evaluating each model's proficiency and consistency, we determined that ERNIE Bot 4, ERNIE Bot 3.5, and ChatGPT 4o could advance to the medical cases review section. The medical cases included 100 patients with various categories of diseases, of whom 58 were females (50.6 ± 16.1 years old), and 42 were males (52.5 ± 17.8 years old).

Figure 5 illustrates a case example where three LLMs made diagnoses and generated formulas based on the patient's medical history description. We initially examined whether the diagnoses provided by the LLMs aligned with those of the doctors. For the SEs of disease location, ChatGPT 4o achieved a DSC of 0.74 (95% CI: 0.69–0.78), significantly superior to ERNIE Bot 3.5, which had a DSC of 0.63 (95% CI: 0.59–0.67). There was no significant difference between ChatGPT 4o and ERNIE Bot 4, which recorded a DSC of 0.70 (95% CI: 0.67–0.74). Regarding the SEs of disease nature, ChatGPT 4o again excelled, with a DSC of 0.73 (95% CI: 0.68–0.77), surpassing both ERNIE Bot 3.5 and ERNIE Bot 4. Table 2 and Figure 6 below offer additional details on the DSC of the three models.

**Table 2. Dice Similarity Coefficient and the 95% CI for syndrome elements**

| | ChatGPT 4o | ERNIE 3.5 | ERNIE 4 |
|---|---|---|---|
| **Disease Location** | 0.74 (0.69, 0.78) | 0.63 (0.59, 0.67) | 0.70 (0.67, 0.74) |
| **Disease Nature** | 0.73 (0.68, 0.77) | 0.64 (0.59, 0.69) | 0.64 (0.60, 0.69) |

Numbers in parentheses are 95% CIs.

Following the diagnosis, we shifted our focus to the prescriptions recommended by LLMs. Each prescription consisted of several Chinese herbal medicines and their corresponding dosages. When scoring the prescriptions, the evaluator primarily concentrated on the composition rather than the dosages. In the case example, the three large language models produced distinct prescriptions: ChatGPT 4o recommended a modified version of Yi Guan Jian, ERNIE 3.5 suggested a modified Longdan Xie Gan Tang, and ERNIE 4 proposed a customized herbal formula. Compared to the doctor's prescription, which was also a modified Yi Guan Jian, ChatGPT 4o received 2 points, whereas ERNIE 3.5 and ERNIE 4 received 0 points. According to the scoring criteria outlined previously, ChatGPT 4o achieved the highest score of 119, outperforming ERNIE 3.5 and ERNIE 4. A Cumulative Score Plot in Figure 7 illustrates the pairwise comparisons among the three models. Although ChatGPT 4o appeared to exhibit optimal results, only 45% of its prescriptions received a score of 2. In the meantime, prescriptions with 1 point accounted for the highest proportion in ERNIE 3.5 (49%) and ERNIE 4 (50%). Results are presented in Table 3.

**Table 3. The score distribution of ChatGPT 4o, ERNIE 3.5, and ERNIE 4 in the prescription generation review**

| **Score** | Case Numbers | | |
|---|---|---|---|
| | ERNIE 4 | ChatGPT 4o | ERNIE 3.5 |
| **0** | 31 | 26 | 29 |
| **1** | 50 | 29 | 49 |
| **2** | 19 | 45 | 22 |

## Discussion

This study comprehensively evaluated LLMs' mastery and application of TCM knowledge alongside their abilities for analysis and decision-making. Our research found that three models—ERNIE 4, ERNIE 3.5, and ChatGPT 4o—outperformed others in TCM knowledge recall, correctly answering more than 70% of the questions with a higher level of consistency. In real clinical cases, these three models could analyze patients' conditions, provide accurate diagnoses, and recommend herbal prescriptions accordingly. While their syndrome differentiation skills reached the level of senior doctors, their ability to prescribe medications remained below that of experienced practitioners.

The TCM Examination section assessed the LLMs' grasp of TCM knowledge and their consistency in answering repeated questions. While models like the ERNIE Bot series and ChatGPT 4o demonstrated high overall accuracy and repeatability, a closer look at their performance across different subjects sheds more light on their strengths and weaknesses. Generally, LLMs scored lower in TCM-related subjects compared to those related to modern medicine. Many struggled particularly with TCM Classics, written in Classical Chinese—a language distinct in structure and style from modern Chinese, and with limited instruction-tuning resources available to enhance understanding(22). Restricted access to specialized corpora narrows the scope of their knowledge, resulting in suboptimal performance on domain-specific questions. Therefore, the future development of a more competent LLM in the TCM domain should prioritize the inclusion of high-quality, diverse, and well-annotated Chinese medical texts in the training process.

It is worth mentioning that ChatGPT 4o, developed by OpenAI, was on par with the ERNIE series and surpassed GLM-4 in the TCM examination. For TCM syndrome diagnoses and prescription generation, ChatGPT 4o even outperformed the ERNIE series. This suggests that, while the ERNIE series excels in knowledge retrieval and handles fact-based exam questions effectively, ChatGPT 4o demonstrates a superior ability to analyze complex inputs by simulating reasoning processes akin to human doctors. Its powerful summarization and analytical capabilities(23) make it more adaptable for tasks requiring interpretation rather than straightforward factual recall.

ChatGPT 4o, ERNIE 3.5, and ERNIE 4 had relatively high diagnostic efficacy in the medical case review section. However, their performance in recommending herbal formulas was mediocre. Less than 50% of the formulas generated by the LLMs received 2 points compared to the prescriptions of experienced doctors. This challenge is compounded by the complexity of real-world medical cases and the variability and flexibility in the TCM treatment process. Patient conditions are often complicated in real-world cases, with multiple pathological factors present simultaneously. Experienced doctors take a holistic view of the patient's situation when prescribing formulas, focusing on the primary issue. For instance, they follow principles such as "symptomatic treatment in acute conditions, radical treatment in chronic cases", and flexibly modify prescriptions based on their experience. However, LLMs have not yet reached this level of sophistication. For example, in the previously presented case, where the primary issue for this patient was liver yin deficiency, the doctor prescribed Modified Yi Guan Jian accordingly. ERNIE 3.5 also identified yin deficiency but suggested Longdan Xiegan Tang, which could potentially worsen the patient's condition in this case. Experienced physicians can grasp the root of the medical condition while considering the patient's circumstances and the interactions among herbs, which is a difficult task for LLMs at the current stage. Future TCM-LLMs should be trained using a more comprehensive dataset that includes real-world medical cases and expert annotations explaining the rationale behind treatment strategies. Though LLMs' performance in prescription generation needs improvement, the breadth of their knowledge and the ability to extract and summarize information should not be underestimated. They can augment human intuition by offering alternative perspectives and assist clinicians in exploring a broader range of treatment possibilities, leading to more informed and well-rounded decisions.

There were limitations to the study. First, due to ethical and safety concerns(24), we cannot directly apply the formulas prescribed by LLMs to patients, making analyzing their effectiveness in clinical practice challenging. As a compromise, we compare the models' diagnoses and prescribed formulas with those of physicians. Nevertheless, there are various schools of thought in TCM, which lead to significant variability in the formulas prescribed by different doctors. Using the doctor's prescription as a reference point, our scoring criteria may introduce bias when evaluating the prescription-generating abilities of large language models. Besides, our study had only one evaluator, which inevitably allowed personal experience and subjective judgment to influence the results. Selecting more representative cases, involving multiple experienced doctors in the assessment process, and refining the evaluation criteria to account for additional dimensions, such as the formula's rationality and safety, will improve the study's reliability. Third, when evaluating the prescriptions recommended by LLMs, we only focused on the similarity of the herbal combinations and overlooked the dosage, which was also a crucial factor in determining the effectiveness of a formula. Future research should address this limitation by incorporating a comparison of not only the prescribed herbs but also their respective dosages.

## Conclusion

In conclusion, while LLMs demonstrated strong proficiency in TCM examination tasks and performed commendably in diagnostic scenarios, they faced challenges in generating accurate and clinically effective herbal prescriptions. Future research should focus on enhancing the practicality and safety of LLMs by incorporating more diverse training data and refining their understanding of TCM treatment principles. The potential applications of LLMs in clinical practice and education within the TCM domain hold significant promise and warrant further exploration.

## Data sharing statement

Data supporting the results demonstrated by this study are available within the main text and the Supplementary Information. Other raw data is not shared due to data protection issues. For further details, the corresponding author may be contacted (sunbaog@mail.sysu.edu.cn).

## Declaration of interests

The authors declare no competing interests.

## Acknowledgments

This study was supported by the National Nature Science Foundation of China (NO: 82205248 and NO: 82272076) and Guangdong Natural Science Foundation (NO：2022A1515012133).

## Contributions

Xiao-Zhou Lu, Hang-Tong Hu and Ting Xiang contributed equally to this article. Ze-Xiong Chen, Bao-Guo Sun, and Wei Wang supervised the study. Xiao-Zhou Lu, Wei Wang, and Bao-Guo Sun had the idea and designed the study. Jin-Zhen Wu, Peng Liu, and Wei-Ming Ji arranged the question banks creation and data collection. Xiao-Zhou Lu and Ting Xiang did the statistical analysis. Xiao-Zhou Lu and Ting Xiang wrote the draft report. Hang-Tong Hu and Wei Wang performed critical revisions on the manuscript. Guan-Heng He, Ze-Xiong Chen, and Bao-Guo Sun helped to organize and conduct the study. All authors contributed to the analysis and interpretation of data. All authors revised the report and approved the final version before submission.

**Reference**

1. Cheung F. TCM: Made in China. Nature. 2011;480(7378):S82-3. doi: 10.1038/480S82a

2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei JS, Chung HW, et al. Large language models encode clinical knowledge (vol 620, pg 172, 2023). Nature. 2023. doi: 10.1038/s41586-023-06455-0

3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature Medicine. 2023;29(8):1930-40. doi: 10.1038/s41591-023-02448-8

4. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS digital health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198

5. Schubert MC, Wick W. Performance of large language models on a neurology board-style examination (vol 6, e2346721, 2023). Jama Network Open. 2024;7(1). doi: 10.1001/jamanetworkopen.2024.0194

6. Bhayana R, Nanda B, Dehkharghanian T, Deng Y, Bhambra N, Elias G, et al. Large Language Models for Automated Synoptic Reports and Resectability Categorization in Pancreatic Cancer. 2024;311(3):e233117. doi: 10.1148/radiol.233117

7. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for

Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. 2023;307(4):e230725. doi: 10.1148/radiol.230725

8. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. 2024;1(1):AIp2300031. doi: doi:10.1056/AIp2300031

9. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. Jama Internal Medicine. 2023;183(6):589-96. doi: 10.1001/jamainternmed.2023.1838

10. Xu WX, Wang L, Zhang MC, Zhu JL, Yan JQ, Wu QT. A joint entity Relation Extraction method for document level Traditional Chinese Medicine texts. Artificial Intelligence in Medicine. 2024;154. doi: 10.1016/j.artmed.2024.102915

11. AlSaad R, Abd-alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal Large Language Models in Health Care:Applications,Challenges, and Future Outlook. Journal of Medical Internet Research. 2024;26. doi: 10.2196/59505

12. Zhang H, Wang X, Meng Z, Jia Y, Xu D. Qibo: A Large Language Model for Traditional Chinese Medicine. doi:

13. Zhu J, Gong Q, Zhou C, Luan H. ZhongJing: A Locally Deployed Large Language Model for Traditional Chinese Medicine and Corresponding Evaluation Methodology: A Large Language Model for data fine-tuning in the field of Traditional Chinese Medicine, and a new evaluation method called TCMEval are proposed. ISAIMS '23: Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science. 2023:1036-42. doi: 10.1145/3644116.3644294

14. Yang G, Shi J, Wang Z, Liu X, Wang G. TCM-GPT: Efficient Pre-training of Large Language Models for Domain Adaptation in Traditional Chinese Medicine arXiv. doi:

15. Hua R, Dong X, Wei Y, Shu ZX, Yang PC, Hu YH, et al. Lingdan: enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models. Journal of the American Medical Informatics Association. 2024. doi: 10.1093/jamia/ocae087

16. Tan Y, Zhang Z, Li M, Pan F, Duan H, Huang Z, et al. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. Computers in biology and medicine. 2024;172:108290. doi: 10.1016/j.compbiomed.2024.108290

17. Yizhen L, Shaohan H, Jiaxing Q, Lei Q, Dongran H, Zhongzhi L. Exploring the Comprehension of ChatGPT in Traditional Chinese Medicine Knowledge. doi:

18. Zhu LX, Mou WM, Lai YC, Lin JD, Luo P. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on Traditional Chinese Medicine highlights the need for localized models. Journal of Translational Medicine. 2024;22(1). doi: 10.1186/s12967-024-05128-4

19. Xu F, Cui WQ, Kong Q, Tang ZH, Dong JC. A Real-World Evidence Study for Distribution of Traditional Chinese Medicine Syndrome and Its Elements on Respiratory Disease. Evidence-Based Complementary and Alternative Medicine. 2018;2018. doi: 10.1155/2018/8305892

20. Wei JP, Wu R, Zhao DD. Analysis on Traditional Chinese Medicine syndrome elements and relevant factors for senile diabetes. Journal of Traditional Chinese Medicine. 2013;33(4):473-8. doi: 10.1016/s0254-6272(13)60151-x

21. Khalafi S, Raissi S, editors. Comparison of Jaccard , Dice , Cosine Similarity Coefficient To Find Best Fitness2013.

22. Rajendran P, Yang Y, Niedermayr TR, Gensheimer M, Beadle B, Le Q-T, et al. Large Language Model-Augmented Auto-Delineation of Treatment Target Volume in Radiation Therapy. arXiv (USA).

2024. doi:

23. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. Journal of Medical Internet Research. 2024;26. doi: 10.2196/56110

24. Fournier-Tombs E, McHardy J. A Medical Ethics Framework for Conversational Artificial Intelligence. Journal of Medical Internet Research. 2023;25. doi: 10.2196/43068

**Supplement 1. Overview of the eight LLMs' characteristics**

| Model | Release Date | Developed by | Model Parameters | Key Features |
|---|---|---|---|---|
| **ChatGPT 3.5** | 2022 | OpenAI | 175B | Excellent language understanding; Supports multi-turn conversations. |
| **ChatGPT 4o** | 2024 | OpenAI | 1T+ | Improved reasoning and language generation quality. |
| **Gemini** | 2023 | Google DeepMind | N/A | Focuses on multimodal tasks. |
| **Gemini Advanced** | 2024 | Google DeepMind | N/A | Enhanced reasoning and complex task processing. |
| **Ernie Bot 3.5** | 2023 | Baidu | 100B | Excels in Chinese language understanding. |
| **Ernie Bot 4** | 2023 | Baidu | 1T | Stronger multitask learning and reasoning, optimized for Chinese. |
| **ChatGLM 3** | 2023 | Tsinghua University | N/A | Innovative architecture and improved contextual understanding. |
| **ChatGLM 4** | 2024 | Tsinghua University | N/A | Enhanced language generation quality and improved reasoning. |

Abbreviations: B: Billions; T: Trillions

N/A: Training parameters data are unknown.

**supplement 1 A sample of the scoring sheet for medical cases review**

| Case No. 1 | Male, 75 years old, with chronic HBV infection and suspected intrahepatic space-occupying lesion. The patient reports abdominal distension, poor appetite, no dry mouth or bitter taste, yellow urine, loose stools, and stable body weight. The tongue is slightly red with a thin white coating, and the pulse is thin and slippery. | | | |
|---|---|---|---|---|
| | Doctor | Model 1 | Model 2 | Model 3 |
| Prescription | Bupleurum root (柴胡) : 10g<br>Scutellaria root (黄芩) : 15g<br>Pinellia tuber (半夏) : 10g<br>Codonopsis root (党参) : 30g<br>Jujube fruit (大枣) : 15g<br>Licorice root (甘草) : 5g<br>Barbed Skullca (半枝莲) : 30g<br>Scorpion (全蝎) : 5g<br>Barley sprout (麦芽) : 30g<br>Turmeric tuber (郁金) : 15g<br>Chicken gizzard lining (鸡内金) : 15g<br>Dried ginger (干姜) : 5g | Bupleurum root (柴胡) : 10g<br>White Peony root (白芍) : 15g<br>Bitter Orange (枳壳) : 10g<br>Licorice root (甘草) : 6g<br>Codonopsis root (党参) : 15g<br>Poria (茯苓) : 20g<br>Atractylodes rhizome (白术): 15g<br>Virgate Wormwood (茵陈): 20g<br>Gardenia fruit (栀子): 10g<br>Scutellaria root (黄芩): 10g | Bupleurum root (柴胡) : 10g<br>White Peony root (白芍) : 12g<br>Bitter Orange (枳壳) : 10g<br>Cyperus rhizome (香附) : 10g<br>Virgate Wormwood (茵陈) : 15g<br>Gardenia fruit (栀子) : 10g<br>Scutellaria root (黄芩) : 10g<br>Atractylodes rhizome (苍术) : 10g<br>Poria (茯苓) : 15g<br>Licorice root (甘草) : 6g<br>Atractylodes rhizome (白术) : 12g<br>Alisma rhizome (泽泻) : 10g<br>Plantain seed (车前子) : 10g<br>Salvia root (丹参) : 15g | Bupleurum root (柴胡) : 10g<br>Angelica root (当归) : 10g<br>White Peony root (白芍) : 15g<br>Atractylodes rhizome (白术) : 15g<br>Poria (茯苓) : 15g<br>Licorice root (甘草) : 6g<br>Mint (薄荷) : 6g<br>Codonopsis root (党参) : 15g<br>Coix seed (薏苡仁) : 20g<br>Amomum fruit (砂仁) : 6g<br>Lotus seed (莲子肉) : 10g<br>Platycodon root (桔梗) : 6g |
| Score | | | | |

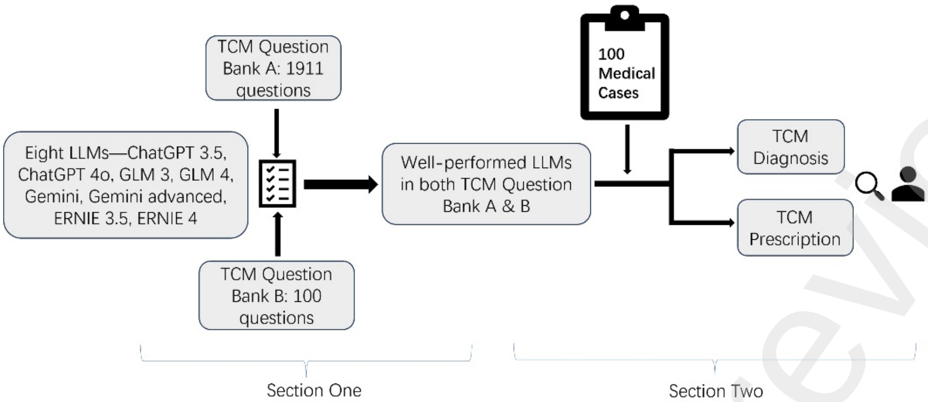
0 = do not match,

1 = partially matches

2 = almost fully match

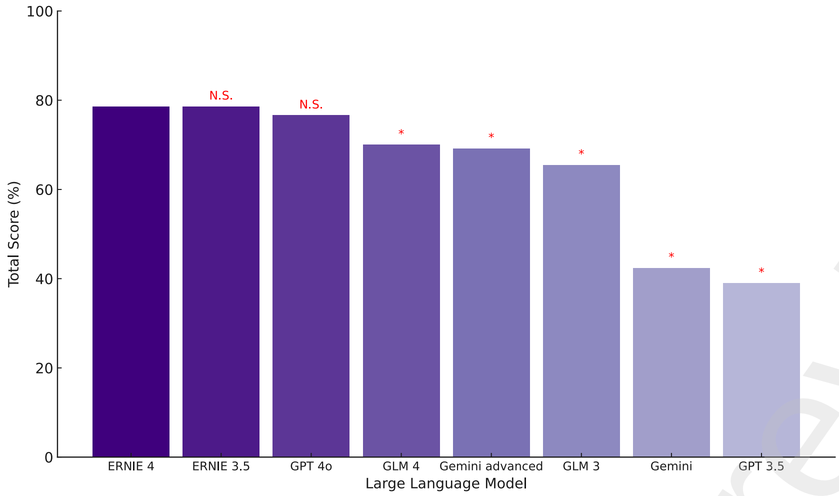
## Figures

**Figure 1: Research workflow diagram.**



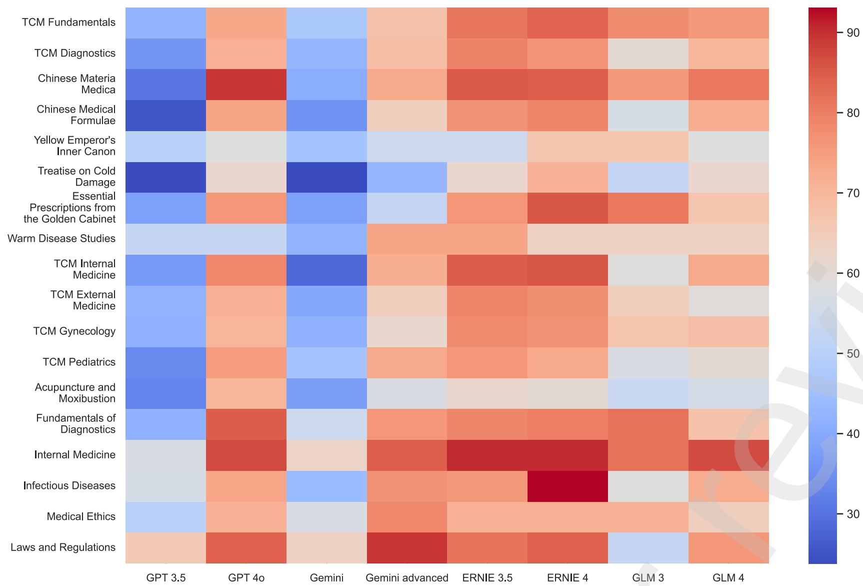

Section One: LLMs' performance in TCM examination. Section Two: LLMs' performance in TCM clinical settings.

**Figure 2: Performance of LLMs in the *TCM Question Bank A*.**



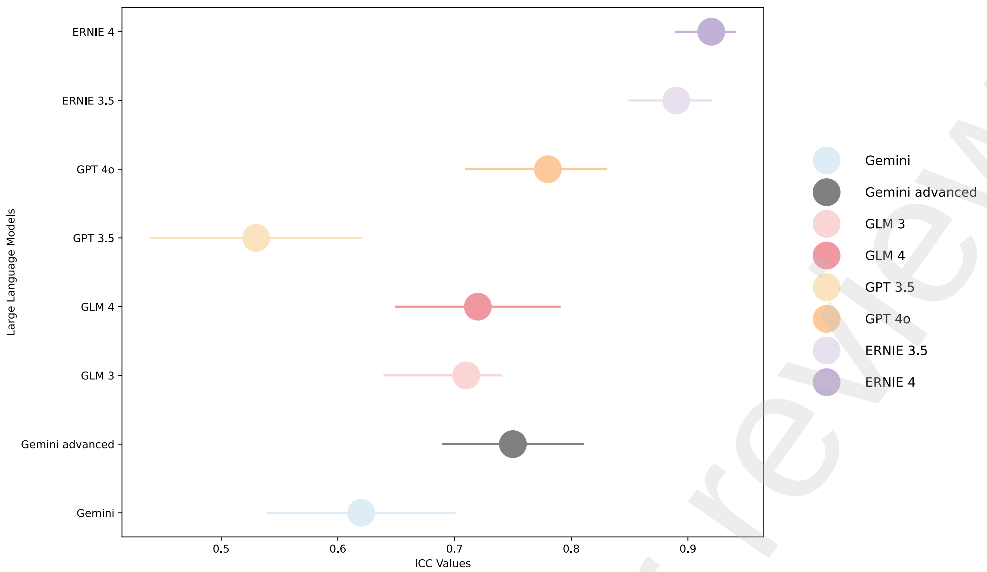

N.S. represents Not Statistically Significant; * indicates $p < .05$

**Figure 3: LLMs' performance across different subjects.**




The color gradient represents the score distribution, with warmer colors indicating higher scores and cooler colors indicating lower scores. Each cell corresponds to a specific model's performance in a given subject.

**Figure 4: ICC values of LLMs on the TCM Question Bank B**




The lollipop chart illustrates the ICC values of various LLMs. Each lollipop circle represents the ICC value for a specific model, with the line extending to show the confidence interval.

**Figure 5: A case example of the diagnosis and formulas generated by ChatGPT 4o, ERNIE 3.5, and ERNIE 4.**

A 42-year-old male reported there was occasional pain under his right rib. His mouth was dry in the morning, and his stool was loose. An ultrasound examination showed he had liver cysts and gallbladder polyps with normal liver function. His tongue was red with little coating, and his pulse was thin and slippery.

**Doctor**

**Syndrome Element Analysis:**
**Pathological Location**: Liver, Gallbladder, Spleen
**Pathological Nature**: Qi stagnation, yin deficiency, and dampness.

**Recommended Formula: Modified Yi Guan Jian（一贯煎加减）**
**Ingredients and Dosages:** Northern Sand Root (北沙参) 15g, Ophiopogon (麦冬) 15g, Sichuan Chinaberry Fruit (川楝子): 10g， Rehmannia Root (生地黄): 10g， Goji Berry (枸杞子): 15g， Angelica (当归): 10g， Turmeric (郁金): 15g， Aucklandia Root (木香): 10g， White Peony Root (白芍): 15g， Mume Fruit (乌梅): 15g， Breaking Stone (穿破石): 15g， Honey-fried Licorice Root (炙甘草): 5g

ChatGPT 4o： Based on the information provided, I will conduct syndrome differentiation and prescribe a formula for this 42-year-old male patient.

**Syndrome Element Analysis:**
**Pathological Location**: Liver and Gallbladder (肝胆)
**Pathological Nature**: Qi stagnation, yin deficiency, and dampness.

**Recommended Formula: Modified Yi Guan Jian（一贯煎加减）**
**Ingredients and Dosages:** Raw Rehmannia Root (生地黄) 12g, Goji Berry (枸杞子) 9g, Northern Sand Root (沙参) 9g, Ophiopogon (麦冬) 9g, Angelica (当归) 9g, Sichuan Chinaberry Fruit (川楝子) 6g, Poria (茯苓) 9g, Alisma (泽泻) 9g, White Atractylodes (白术) 9g.

ERNIE Bot 3.5： Based on the patient's symptoms and examination results, a syndrome differentiation diagnosis in Traditional Chinese Medicine (TCM) can be made as follows:

**Syndrome Differentiation Diagnosis**
**Pathological Location**: Liver, Gallbladder, and Stomach
**Nature Elements**: Heat, Deficiency (Yin Deficiency), and Dampness.

**Formula: Modified Long Dan Xie Gan Tang (龙胆泻肝汤) combined with Yin Chen Hao Tang (茵陈蒿汤) and Liu Wei Di Huang Wan (六味地黄丸).**
**Ingredients and Dosages:** Gentianae Scabrae Radix (龙胆) 9g, Paeoniae Radix Alba (白芍) 9g, Artemisiae Scopariae Herba (茵陈) 15g, Poria Cocos (茯苓) 12g, Rehmanniae Radix Preparata (熟地黄)12g, Corni Fructus (山茱萸) 9g, Dioscoreae Rhizoma (山药) 12g, Alismatis Rhizoma (泽泻) 9g, Glycyrrhizae Radix Preparata (炙甘草) 6g

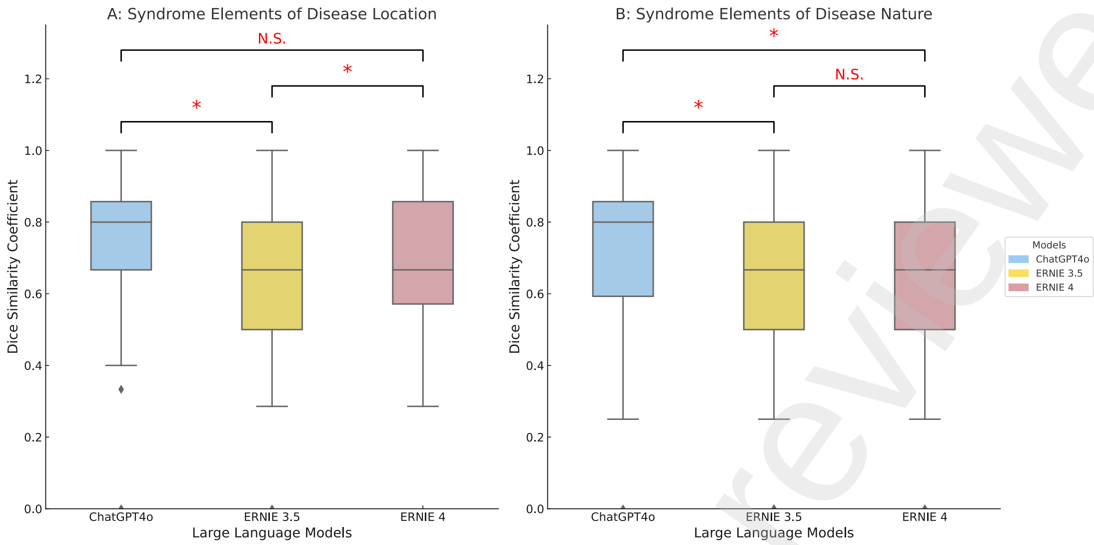
ERNIE Bot 4.0： Based on the symptoms and signs described, the patient's condition can be analyzed using the principles of Traditional Chinese Medicine (TCM).

**Syndrome Differentiation Diagnosis**
**Pathological Location**: Liver, Gallbladder, and Spleen
**Nature Elements**: Heat, Qi Deficiency, and Dampness

**Herbal Formula: A customized herbal formula**
**Ingredients and Dosages:** Radix Bupleuri (柴胡) 9g, Rhizoma Atractylodis Macrocephalae (白术)9g, Radix Scutellariae(黄芩) 9g, Fructus Aurantii Immaturus(枳壳) 6g, Sclerotium Poriae Cocos (茯苓) 9g, Radix Glycyrrhizae(甘草) 3g.

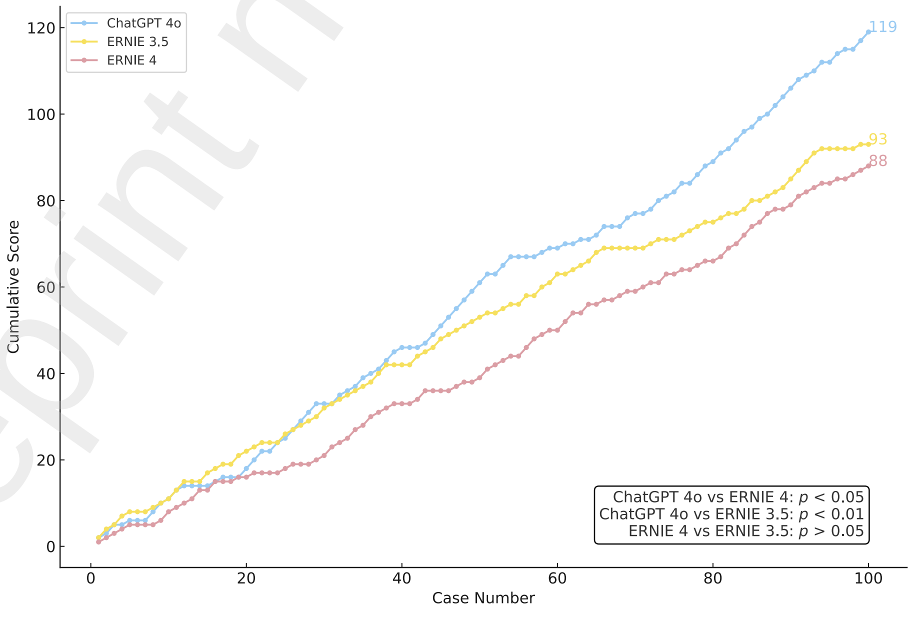
A clinical case is presented, including the diagnoses by the doctor and the LLMs (highlighted with an orange background) and their respective prescriptions (highlighted with a light blue background).

**Figure 6: Comparison of three LLMs' DSC in diagnosing disease location and nature.**

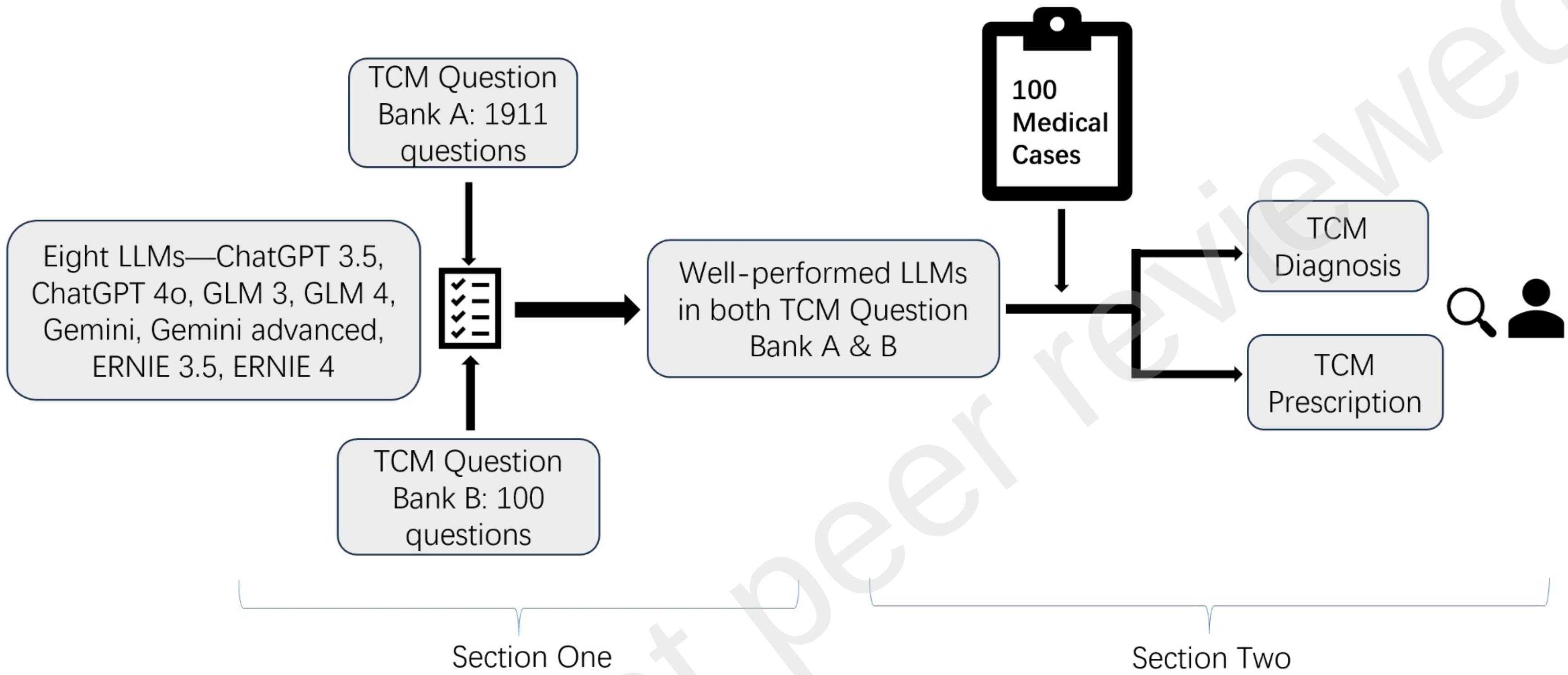


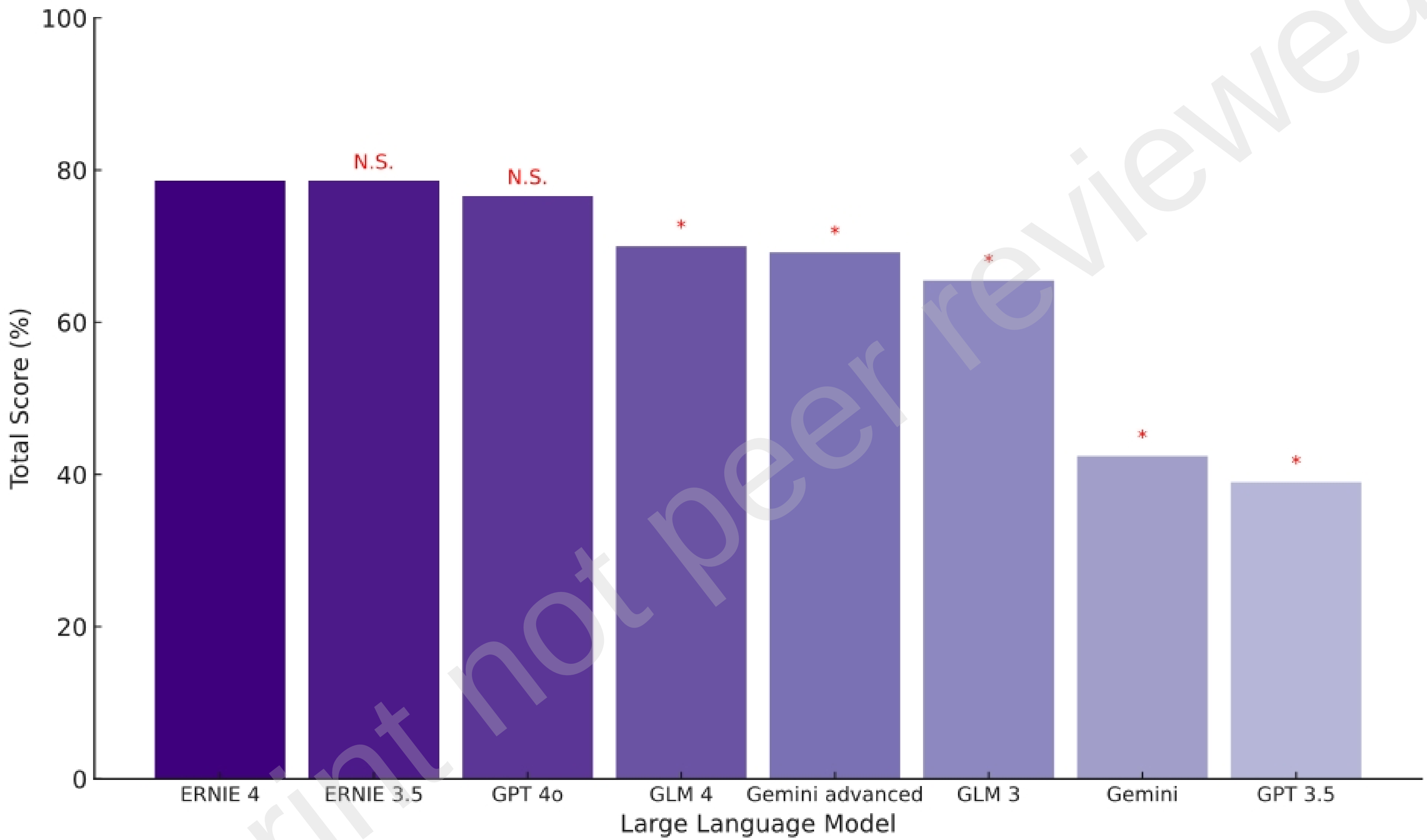

In the box plot, the boxes represent the interquartile range (IQR), with their edges indicating Q1 (the 25th percentile) and Q3 (the 75th percentile). The DSCs of the three models are compared with each other, where N.S. indicates Not Statistically Significant, and * denotes $p < .05$.

**Figure 7: Pairwise Comparisons of prescription generation performance among ChatGPT 4o, ERNIE 3.5, and ERNIE 4.**

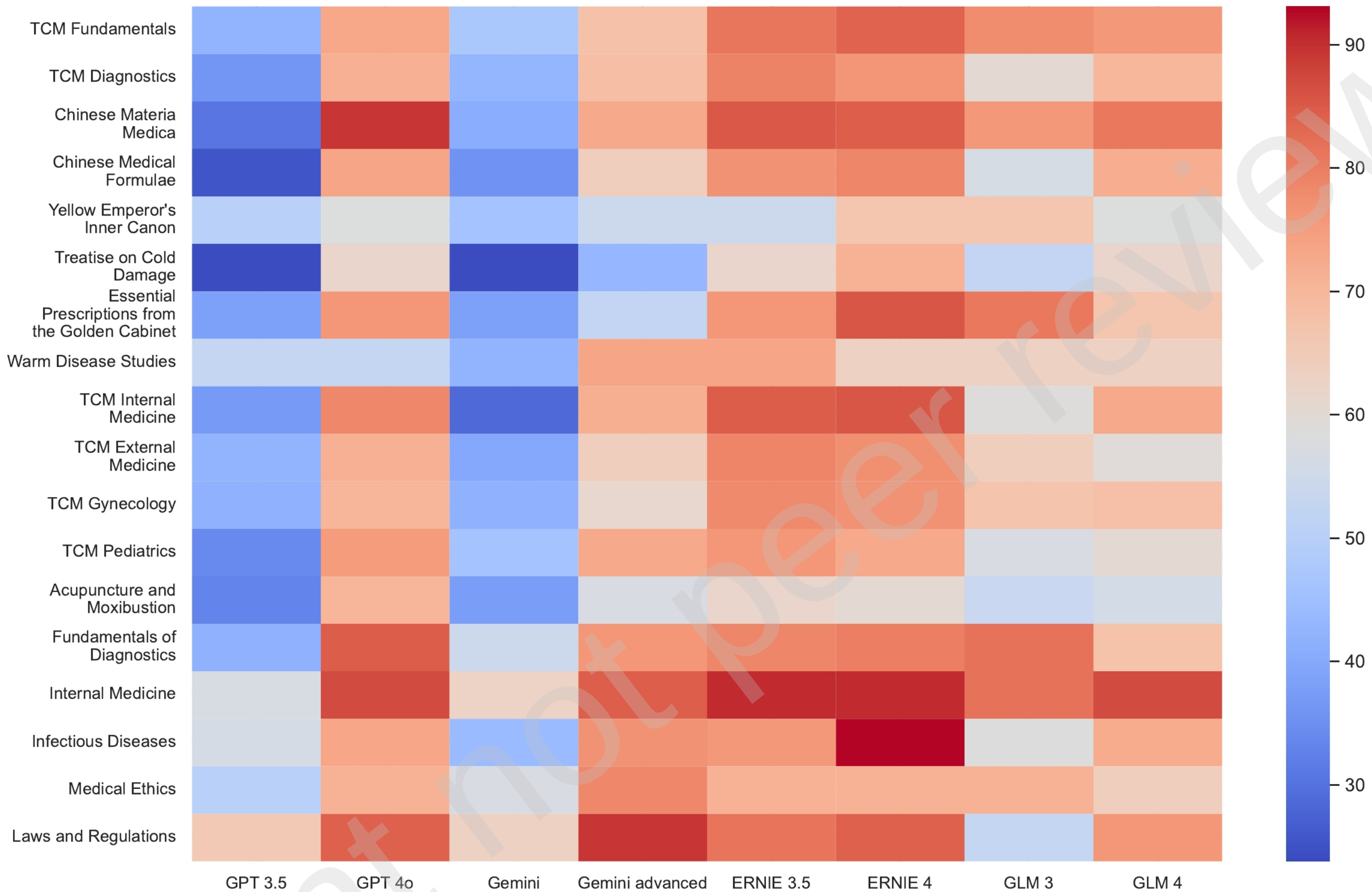


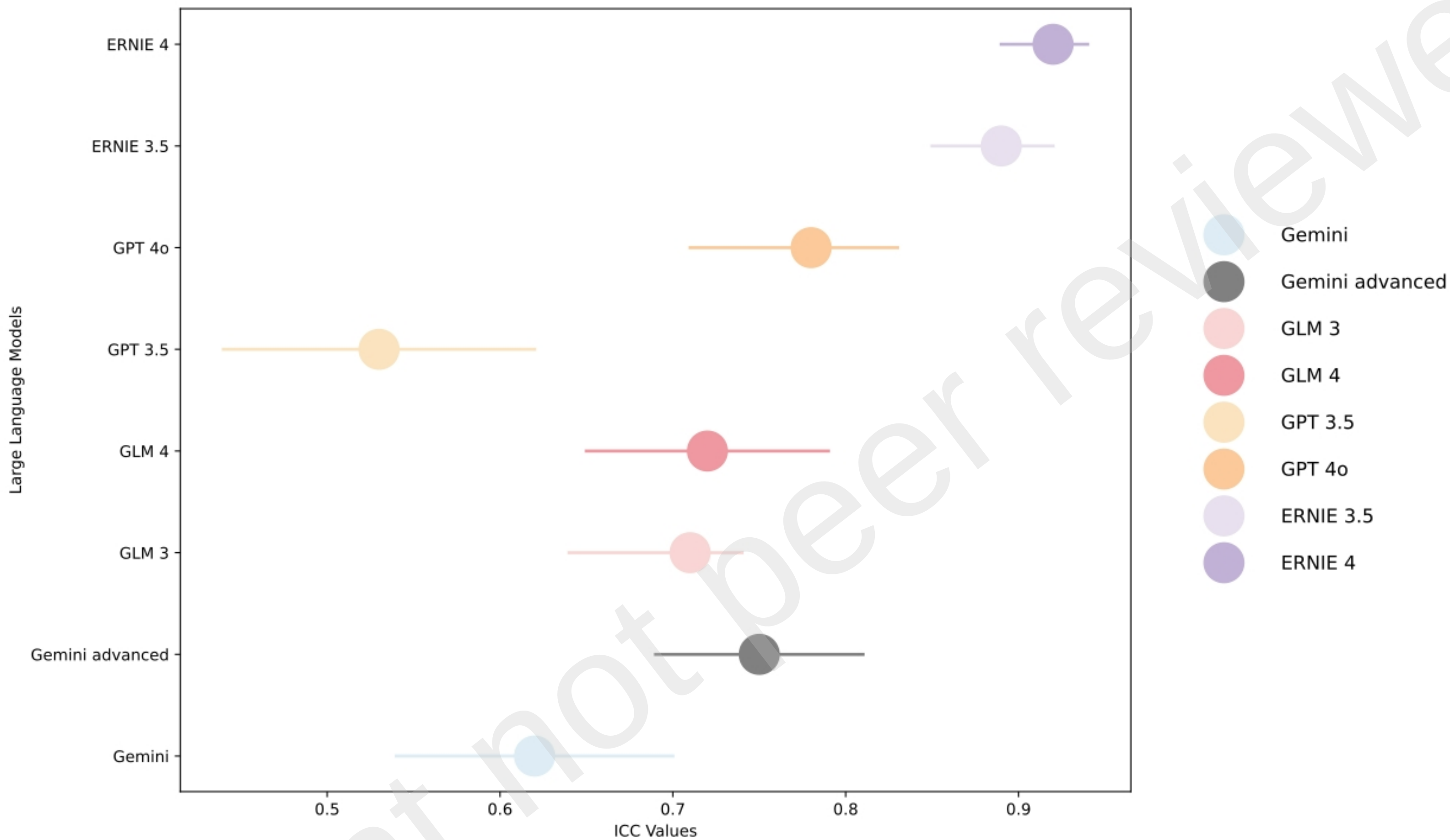
The plot illustrates differences in how effectively ChatGPT 4o, ERNIE 3.5, and ERNIE 4 generate prescriptions; each curve represents one model, with higher curves indicating better overall performance.


TCM Question Bank A: 1911 questions
100 Medical Cases
Eight LLMs—ChatGPT 3.5, ChatGPT 4o, GLM 3, GLM 4, Gemini, Gemini advanced, ERNIE 3.5, ERNIE 4
Well-performed LLMs in both TCM Question Bank A & B
TCM Diagnosis
TCM Prescription
TCM Question Bank B: 100 questions
Section One
Section Two



100
80
60
40
20
0
Total Score (%)
N.S.
N.S.
*
*
*
*
*
ERNIE 4
ERNIE 3.5
GPT 4o
GLM 4
Gemini advanced
GLM 3
Gemini
GPT 3.5
Large Language Model


TCM Fundamentals
TCM Diagnostics
Chinese Materia Medica
Chinese Medical Formulae
Yellow Emperor's Inner Canon
Treatise on Cold Damage
Essential Prescriptions from the Golden Cabinet
Warm Disease Studies
TCM Internal Medicine
TCM External Medicine
TCM Gynecology
TCM Pediatrics
Acupuncture and Moxibustion
Fundamentals of Diagnostics
Internal Medicine
Infectious Diseases
Medical Ethics
Laws and Regulations
GPT 3.5
GPT 4o
Gemini
Gemini advanced
ERNIE 3.5
ERNIE 4
GLM 3
GLM 4
90
80
70
60
50
40
30



ERNIE 4
ERNIE 3.5
GPT 4o
GPT 3.5
GLM 4
GLM 3
Gemini advanced
Gemini
Large Language Models
0.5
0.6
0.7
0.8
0.9
ICC Values
Gemini
Gemini advanced
GLM 3
GLM 4
GPT 3.5
GPT 4o
ERNIE 3.5
ERNIE 4

A 42-year-old male reported there was occasional pain under his right rib. His mouth was dry in the morning, and his stool was loose. An ultrasound examination showed he had liver cysts and gallbladder polyps with normal liver function. His tongue was red with little coating, and his pulse was thin and slippery.

**Doctor**

**Syndrome Element Analysis:**
**Pathological Location**: Liver, Gallbladder, Spleen
**Pathological Nature**: Qi stagnation, yin deficiency, and dampness.

**Recommended Formula: Modified Yi Guan Jian（一贯煎加减）**
**Ingredients and Dosages:** Northern Sand Root (北沙参) 15g, Ophiopogon (麦冬) 15g, Sichuan Chinaberry Fruit (川楝子): 10g, Rehmannia Root (生地黄): 10g, Goji Berry (枸杞子): 15g, Angelica (当归): 10g, Turmeric (郁金): 15g, Aucklandia Root (木香): 10g, White Peony Root (白芍): 15g, Mume Fruit (乌梅): 15g, Breaking Stone (穿破石): 15g, Honey-fried Licorice Root (炙甘草): 5g

**ChatGPT 4o：** Based on the information provided, I will conduct syndrome differentiation and prescribe a formula for this 42-year-old male patient.

**Syndrome Element Analysis:**
**Pathological Location**: Liver and Gallbladder (肝胆)
**Pathological Nature**: Qi stagnation, yin deficiency, and dampness.

**Recommended Formula: Modified Yi Guan Jian（一贯煎加减）**
**Ingredients and Dosages:** Raw Rehmannia Root (生地黄) 12g, Goji Berry (枸杞子) 9g, Northern Sand Root (沙参) 9g, Ophiopogon (麦冬) 9g, Angelica (当归) 9g, Sichuan Chinaberry Fruit (川楝子) 6g, Poria (茯苓) 9g, Alisma (泽泻) 9g, White Atractylodes (白术) 9g.

**ERNIE Bot 3.5：** Based on the patient's symptoms and examination results, a syndrome differentiation diagnosis in Traditional Chinese Medicine (TCM) can be made as follows:

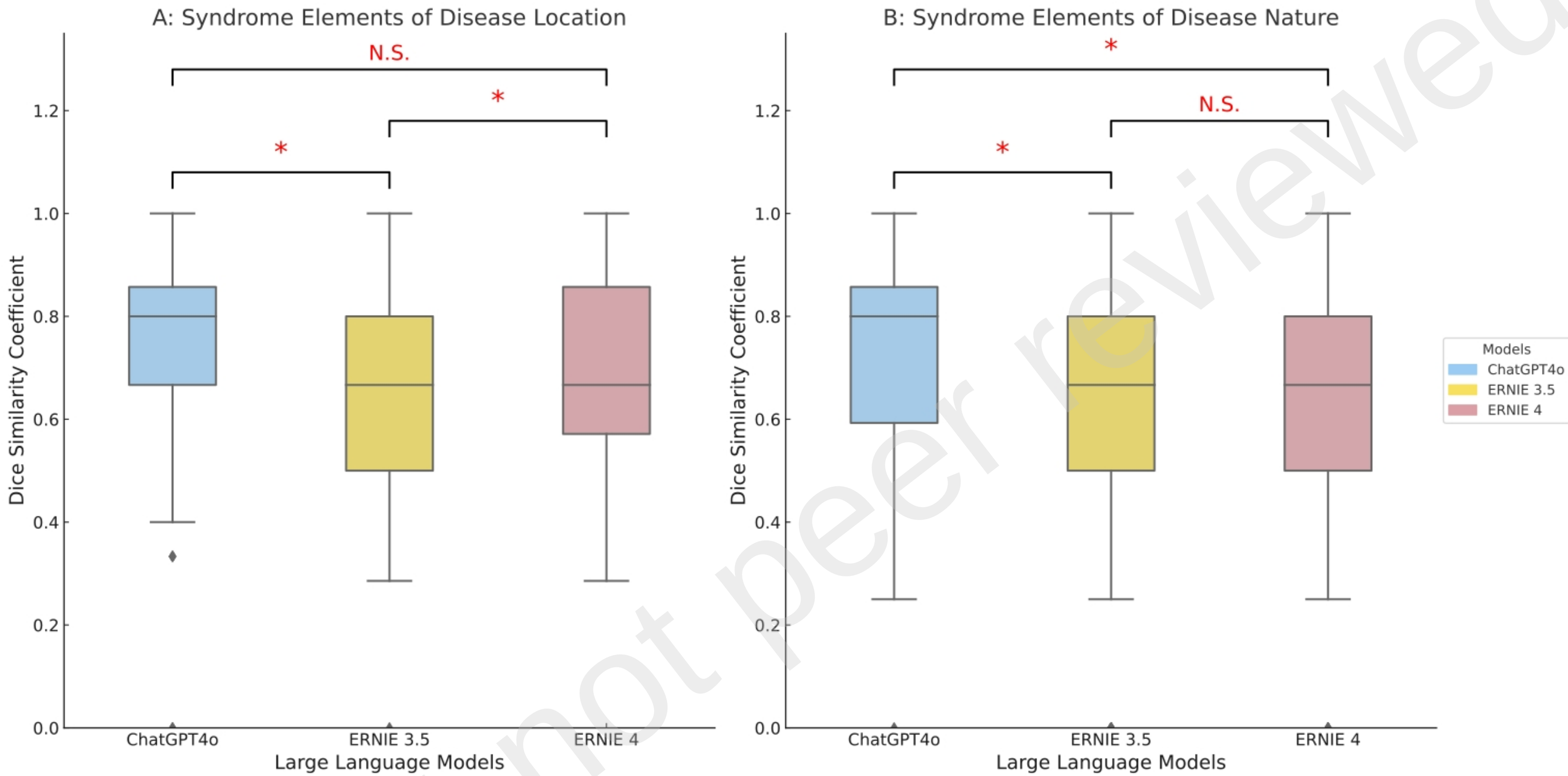
**Syndrome Differentiation Diagnosis**
**Pathological Location**: Liver, Gallbladder, and Stomach
**Nature Elements**: Heat, Deficiency (Yin Deficiency), and Dampness.

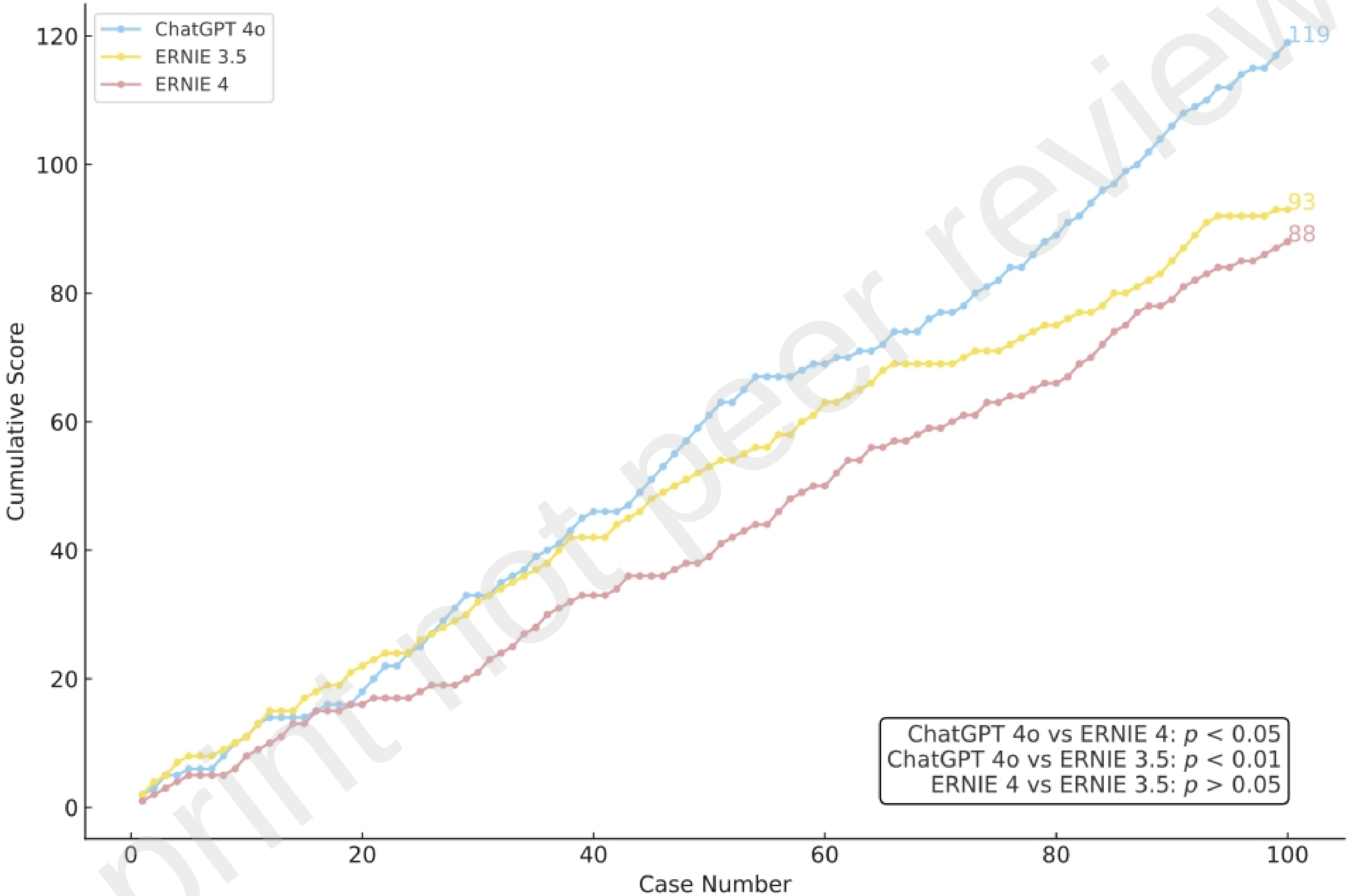
**Formula**: **Modified Long Dan Xie Gan Tang (龙胆泻肝汤) combined with Yin Chen Hao Tang (茵陈蒿汤) and Liu Wei Di Huang Wan (六味地黄丸).**
**Ingredients and Dosages**: Gentianae Scabrae Radix (龙胆) 9g, Paeoniae Radix Alba (白芍) 9g, Artemisiae Scopariae Herba (茵陈) 15g, Poria Cocos (茯苓) 12g, Rehmanniae Radix Preparata (熟地黄)12g, Corni Fructus (山茱萸) 9g, Dioscoreae Rhizoma (山药) 12g, Alismatis Rhizoma (泽泻) 9g, Glycyrrhizae Radix Preparata (炙甘草) 6g

**ERNIE Bot 4.0：** Based on the symptoms and signs described, the patient's condition can be analyzed using the principles of Traditional Chinese Medicine (TCM).

**Syndrome Differentiation Diagnosis**
**Pathological Location**: Liver, Gallbladder, and Spleen
**Nature Elements**: Heat, Qi Deficiency, and Dampness

**Herbal Formula: A customized herbal formula**
**Ingredients and Dosages:** Radix Bupleuri (柴胡) 9g, Rhizoma Atractylodis Macrocephalae (白术)9g, Radix Scutellariae(黄芩) 9g, Fructus Aurantii Immaturus(枳壳) 6g, Sclerotium Poriae Cocos (茯苓) 9g, Radix Glycyrrhizae(甘草) 3g.


A: Syndrome Elements of Disease Location
B: Syndrome Elements of Disease Nature
N.S.
*
Dice Similarity Coefficient
Large Language Models
ChatGPT4o
ERNIE 3.5
ERNIE 4
Models
0.0
0.2
0.4
0.6
0.8
1.0
1.2


ChatGPT 4o
ERNIE 3.5
ERNIE 4
119
93
88
Cumulative Score
Case Number
0
20
40
60
80
100
120
ChatGPT 4o vs ERNIE 4: $p < 0.05$
ChatGPT 4o vs ERNIE 3.5: $p < 0.01$
ERNIE 4 vs ERNIE 3.5: $p > 0.05$