

TCMBench: A Comprehensive Benchmark for Evaluating Large Language Models in Traditional Chinese Medicine

Wenjing Yue^{1,2}, Xiaoling Wang^{*2}, Wei Zhu², Ming Guan², Changzhi Sun³, Huanran Zheng², Pengfei Wang², Xin Ma⁴

¹Shanghai Institute of AI for Education, East China Normal University, Shanghai, China

²School of Computer Science and Technology, East China Normal University, Shanghai, China

³Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai, China

⁴Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China

{wjyue, wzhu, mguan, hrzheng, pfwang}@stu.ecnu.edu.cn, xlwang@cs.ecnu.edu.cn, czsun.cs@gmail.com, nicole.ma@me.com

APPENDIX A

THE QUESTION TYPE OF TCM-EB

There are three types of questions in TCMIE, which we have organized into TCM-EB:

- **The single-sentence best-choice questions(A1) and the case summary best-choice questions(A2) type:** It consists of a question stem and five options with correct one, as shown in Figure 1.
- **The best choice questions for case group(A3) type:** The stem presents a patient-centered case, followed by multiple sub-questions, each offering five options with one correct answer. It primarily centers on clinical applications, as shown in Figure 2.
- **The standard compatibility questions(B1) type:** Multiple sub-questions share the same five options, where each option may be chosen zero, one, or multiple times. There is one correct answer among the five options for each sub-question, as shown in Figure 3.

一、A1/A2型选择题（每一道题下面有A、B、C、D、E五个备选答案，请从中选择一个最佳答案。） 二、A1/A2 type choice question: Each test question includes a question stem and five alternative answers (A, B, C, D, and E). The stem precedes the alternatives, and there is only one correct answer. The stem is presented as a discourse question, either narratively or negatively (A1), or as a simple clinical case (A2).	
Question Requirement	
【问题】患者，女，43岁。入院时诊断为肠痈。现腹皮急，全腹压痛、反跳痛，腹胀，恶心、呕吐，大便不爽，次数增多，小便频数，时时汗出，皮肤甲错。二目下陷，口干而黄，舌红苔黄糙，脉细数。其证候是（ ）。 【Question] A 43-year-old female patient is diagnosed with an intestinal carbuncle upon admission. She presents with abdominal skin contracture, tenderness, rebound pain, abdominal distension, nausea, vomiting, uncomfortable bowel movements, increased frequency of bowel movements, frequent urination, frequent sweating, and skin and nail irregularities. Her eyes are sunken, her mouth is dry with a foul odor, her tongue is red with a yellow and rough coating, and her pulse is threadly and rapid. What the syndrome is ().	
Question	A. 积热不散，热盛肉腐 Accumulated heat does not dissipate, hyperactivity of heat bringing about rottenness of muscle B. 阳明腑实，热盛伤阴 Yangming Fu (viscera) repletion, excessive heat injuring Yin C. 寒湿内蕴，瘀血凝滞 Cold-dampness internal accumulation, stasis of stagnant blood D. 湿热内蕴，气血凝滞 Damp-heat internal accumulation, stagnation of Qi and blood E. 邪毒内蕴，瘀血凝滞 Pathogenic toxins internal accumulation, stasis of stagnant blood
Standard Answer	B
Analysis	【解析】肠痈是由胃肠受损，肠道传化失司，糟粕停滞，气血瘀滞所致。瘀阻日久化热，热盛则肉腐成痈。全腹压痛、反跳痛，腹胀，恶心、呕吐，大便不爽，次数增多为阳明腑实症状；皮肤甲错，二目下陷，口干而黄，舌红苔黄糙，脉细数，为热盛伤阴的临床表现。 Intestinal carbuncle is caused by damage to the stomach and intestines, derangement in the transformation and transportation function of the intestines, stagnation of impurities, Qi stagnation, and blood stasis. Prolonged stasis accumulates heat, and excessive heat can cause the tissues to fester, resulting in an abscess. Symptoms include generalized abdominal tenderness, rebound pain, abdominal distension, nausea, vomiting, constant sweating, uncomfortable bowel movements, and increased frequency point to Yangming Fu repletion. Skin and nail irregularities, sunken eyes, dry and foul mouth, red tongue with a yellow and rough coating, and a threadly and rapid pulse indicate clinical manifestations of excessive heat damaging Yin.
Description of TCM terms: Qi is often described as energy. Yin and Yang refer to the human body's physiological processes and pathological changes.	

Fig. 1. The example of A1/A2 type of questions. The question requirement is indicated in dark blue text, the question along with the five options is in light blue text, the standard answer is in green text, and the standard analysis is in orange text. The related TCM terms are explained in the yellow highlight.

三、A3型题（以下提供若干案例，每个案例下设有若干道试题，请根据案例所提供的信息，在每一道试题下面的A、B、C、D、E五个备选答案中选择一个最佳答案。） Each case is a main test question. Choose the best answer from options A, B, C, D, and E below each question based on the provided case information.	
Question Requirement	
【共用题干】患儿，女，9个月，1个月来患肺炎一直未愈，现干咳少痰，面黄，低热盗汗，口唇干燥，大便燥结，舌红苔黄，指纹紫滞了三天。 The patient is a 9-month-old girl who had pneumonia one month ago, which has not fully resolved. She is currently experiencing a dry cough with minimal phlegm, slight wheezing, low-grade fever, night sweats, dry lips, heat inside, a red and peeled tongue, and the veins in the middle of the palm are congested. The veins on both hands' front edges appear pale purple.	
Sub Question 1	标准答案 1
【解析】患儿患肺炎月余未愈，为小兒肺热咳嗽肺热证。证候：咳嗽少痰，面黄盗汗，手足心热，干咽少痰，面黄肌瘦，口干唇燥，舌红少津，舌少或无苔，脉细数，指纹淡紫。 Based on the clinical presentation of the patient, the diagnosis is "relative pneumonia with cough due to Yin deficiency and lung heat syndrome." The syndrome includes the following symptoms: Persistent cough and wheezing, low-grade fever with night sweats, heat sensation in the hands, feet, and chest, dry cough with minimal phlegm, flushed complexion, dry mouth and constipation, red tongue with reduced moisture and a scanty or peeled coating, rapid and thin pulse, the veins on the palm sides of both hands' front edges appear pale purple.	
Sub Question 2	标准答案 2
【解析】小兒肺热咳嗽肺热证治法：清肺润燥，滋阴止咳。代表方剂：沙参麦冬汤。 The therapy of pediatric pneumonia with cough due to Yin deficiency and lung heat syndrome is nourishing Yin, clearing the lungs, moistening the lungs, and stopping coughing. The represent prescription is Decoction of Glabaria and Ophiopogon.	
Sub Question 3	标准答案 3
【解析】小兒肺热咳嗽肺热证治法：清肺润燥，滋阴止咳。代表方剂：沙参麦冬汤。 The therapy of pediatric pneumonia with cough due to Yin deficiency and lung heat syndrome is nourishing Yin, clearing the lungs, moistening the lungs, and stopping coughing. The represent prescription is Decoction of Glabaria and Ophiopogon.	

Fig. 2. The example of A3 type of questions. The question requirement is indicated in dark blue text, the patient-centered case is in light blue text, the first sub question along with the five options, standard answer and analysis is in green text, the second sub question is in orange text, and the second sub question is in purple text.

三、B1型题（以下提供若干组题，每组题共用在考题前列出的A、B、C、D、E五个备选答案，请从中选择一个与问题最密切的答案，某一个备选答案可能被选择一次、多次或不被选择。） Each question group shares the answers A, B, C, D, and E. Choose the most relevant answer. You can choose an answer once, multiple times, or not.	
Question Requirement	
【共用备选答案】 Shared alternative answers A. 气机不畅 Stagnation of Qi B. 脾虚湿盛 Spleen Deficiency with Dampness Invasion C. 痰饮内停 Drug Possession D. 阴盛火旺 Yin Deficiency with Excessive Fire E. 痰饮内停 Internal retention of phlegm and water-pathogen Blood stasis	
Sub Question 1	标准答案 1
【解析】舌淡红中泛青紫，多因肺气壅滞，或肝郁血瘀，亦可示为先天性心脏病，或某些药物、食物中毒。The tongue is pale red with bluish-purple spots in the center, often due to lung Qi stagnation or liver Qi stagnation with blood stasis, and can also be observed in congenital heart disease or specific drug or food poisoning.	
Sub Question 2	标准答案 2
【解析】舌淡白而有裂纹，多为血虚不润，舌淡白而燥，边有黄褐斑又有裂纹，属脾虚湿盛。A pale tongue with cracks is often due to blood deficiency and lack of moisture. The tongue has a thick, white coating with teeth marks and cracks at the edges. Especially at night, patients feel dry. The clinical performance is spleen deficiency with dampness invasion.	

Fig. 3. The example of B1 type of questions. The question requirement is indicated in dark blue text, five options is in light blue text, the first sub question along with the five options, standard answer and analysis is in green text, and the second sub question is in orange text.

APPENDIX B

THE DISTRIBUTION OF FIVE OPTIONS IN TCM-ED AMONG THE THREE TYPES OF QUESTIONS

We calculate the distribution of the five options across the three types of questions in TCM-ED, as shown in Figure 4. The results indicate that the distribution of choices among the

*Corresponding author.

three question types is relatively uniform.

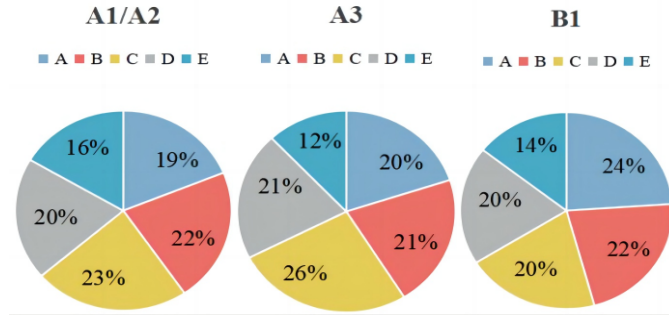


Fig. 4. The distribution of options in TCM-ED among A1/A2 types(left), A3 types(middle), and B1 types(right) of questions.

APPENDIX C THE STATISTICAL INFORMATION OF LLMs.

TABLE I
THE STATISTICAL INFORMATION OF LLMs.

Model	Open Source	Parameters	Domain
GPT-4	×	175B+	General
ChatGPT	×	175B	General
ChatGLM	✓	130B	General
Chinese LLaMa	✓	7B	General
HuaTuo	✓	7B	Chinese Medicine
ZhongJing-TCM	✓	7B	TCM

APPENDIX D PROMPT AND TARGET OUTPUT FORMATS FOR EVALUATION

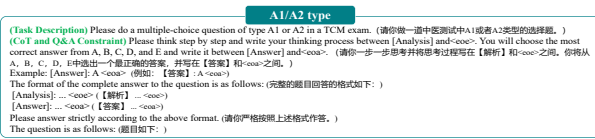


Fig. 5. The zero-shot prompt template target output formats for evaluating A1/A2 type of questions.

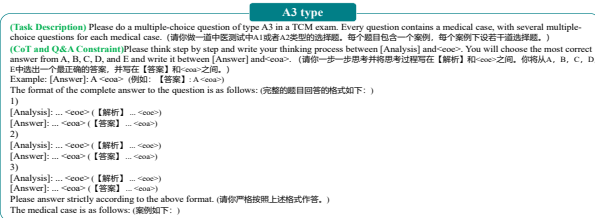


Fig. 6. The zero-shot prompt template target output formats for evaluating A3 type of questions.

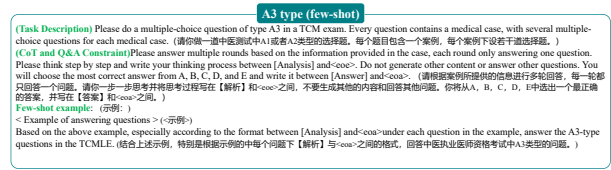


Fig. 7. The few-shot prompt template target output formats for evaluating A3 type of questions.

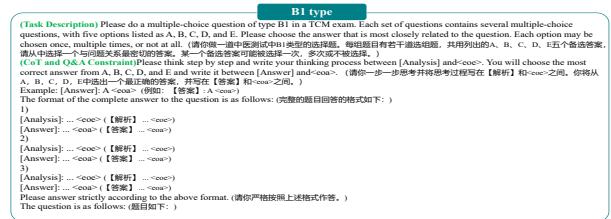


Fig. 8. The zero-shot prompt template target output formats for evaluating B1 type of questions.