

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。入院时诊断为肠痛。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. 寒湿内阻，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】肠痛是由胃肠受损，肠道传化失司，糟粕停滞，气滞血瘀所致。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，舌红无苔，指纹淡紫，诊断为肺炎。患儿双前臂掌侧皮肤出现紫红色斑丘疹，疹面干燥，疹周有红晕，疹间皮肤正常。患儿双前臂掌侧皮肤出现紫红色斑丘疹，疹面干燥，疹周有红晕，疹间皮肤正常。	
Shared question stem (clinical case)	
Sub Question 1	Standard Answer 1
Analysis 1	
Sub Question 2	Standard Answer 2
Analysis 2	
Sub Question 3	Standard Answer 3
Analysis 3	

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（以下提供若干组试题，每组试题共用题干前列出的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	
Shared alternative answers	
Sub Question 1	Standard Answer 1
Analysis 1	
Sub Question 2	Standard Answer 2
Analysis 2	

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.

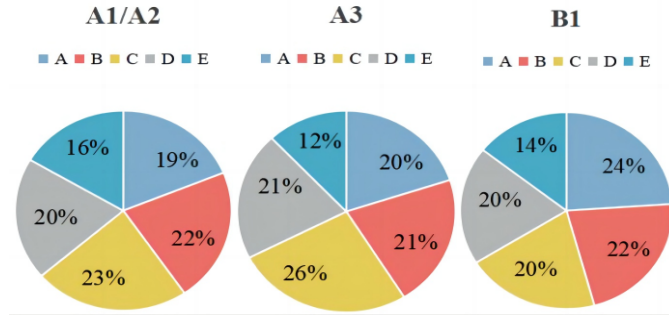


Fig. 4. The distribution of options in TCM-ED among A1/A2 types(left), A3 types(middle), and B1 types(right) of questions.

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 ¹	✓	7B	TCM

APPENDIX D PROMPT AND TARGET OUTPUT FORMATS FOR EVALUATION

A1/A2 type

(Task Description) Please do a multiple-choice question of type A1 or A2 in a TCM exam. (请你做一道中医测试中A1或者A2类型的选择题。)

(CoT and Q&A Constraint) Please think step by step and write your thinking process between [Analysis] and <coe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and <coas>. (请你一步一步思考并将在【解析】和<coe>之间, 你将从A, B, C, D, E中选择一个最正确的答案, 并写在【答案】和<coas>之间。)

Example [Answer]: A <coas> (例如: 【答案】: A <coas>)

The format of the complete answer to the question is as follows: (完整的答案格式如下:)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

Please answer strictly according to the above format. (请你严格按照上述格式作答。)

The question is as follows: (题目如下:)

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

A3 type

(Task Description) Please do a multiple-choice question of type A3 in a TCM exam. Every question contains a medical case, with several multiple-choice questions for each medical case. (请你做一道中医测试中A3类型的选择题。每个题目包含一个案例, 每个案例下设若干道选择题。)

(CoT and Q&A Constraint) Please think step by step and write your thinking process between [Analysis] and <coe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and <coas>. (请你一步一步思考并将在【解析】和<coe>之间, 你将从A, B, C, D, E中选择一个最正确的答案, 并写在【答案】和<coas>之间。)

Example [Answer]: A <coas> (例如: 【答案】: A <coas>)

The format of the complete answer to the question is as follows: (完整的答案格式如下:)

1)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

2)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

3)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

Please answer strictly according to the above format. (请你严格按照上述格式作答。)

The medical case is as follows: (案例如下:)

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

A3 type (few-shot)

(Task Description) Please do a multiple-choice question of type A3 in a TCM exam. Every question contains a medical case, with several multiple-choice questions for each medical case. (请你做一道中医测试中A3类型的选择题。每个题目包含一个案例, 每个案例下设若干道选择题。)

(CoT and Q&A Constraint) Please answer multiple rounds based on the information provided in the case, each round only answering one question. Please think step by step and write your thinking process between [Analysis] and <coe>. Do not generate other content or answer other questions. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and <coas>. (请你根据案例所提供的信息逐步回答问题, 每一轮都只回答一个问题。请你一步一步思考并将在【解析】和<coe>之间, 不要生成其他的内容和回答其他问题。你将从A, B, C, D, E中选择一个最正确的答案, 并写在【答案】和<coas>之间。)

Few-shot example: (示例:)

< Example of answering questions > (示例:)

Based on the above example, especially according to the format between [Analysis] and <coe> under each question in the example, answer the A3-type questions in the TCMLE. (结合上述示例, 特别是根据示例中的每个问题下【解析】与<coe>之间的格式, 回答中医执业医师资格考试中A3类型的问题。)

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

B1 type

(Task Description) Please do a multiple-choice question of type B1 in a TCM exam. Each set of questions contains several multiple-choice questions, with five options listed as A, B, C, D, and E. Please choose the answer that is most closely related to the question. Each option may be chosen once, multiple times, or not at all. (请你做一道中医测试中B1类型的选择题。每道题目前有若干道选择题, 共用列出的A, B, C, D, E五个备选答案, 选项可能选一次、多次或不选。)

(CoT and Q&A Constraint) Please think step by step and write your thinking process between [Analysis] and <coe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and <coas>. (请你一步一步思考并将在【解析】和<coe>之间, 你将从A, B, C, D, E中选择一个最正确的答案, 并写在【答案】和<coas>之间。)

Example [Answer]: A <coas> (例如: 【答案】: A <coas>)

The format of the complete answer to the question is as follows: (完整的答案格式如下:)

1)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

2)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

3)

[Analysis]: ... <coe> (【解析】 ... <coe>)

[Answer]: ... <coas> (【答案】 ... <coas>)

Please answer strictly according to the above format. (请你严格按照上述格式作答。)

The question is as follows: (题目如下:)

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