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

Wenjing Yue[1, 2], Xiaoling Wang*[2], Wei Zhu[2], Ming Guan[2], Changzhi Sun[3], Huanran Zheng[2], Pengfei Wang[2], Xin Ma[4]
[1]*Shanghai Institute of AI for Education, East China Normal University, Shanghai, China*
[2]*School of Computer Science and Technology, East China Normal University, Shanghai, China*
[3]*Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai, China*
[4]*Shanghai 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.

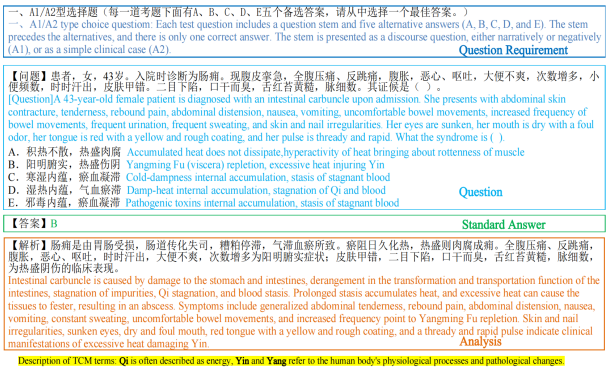

Fig. 1. The example of A1/A2 type of questions. The question requirement is indicated in dark blue text, the question along with the five options is in light blue text, the standard answer is in green text, and the standard analysis is in orange text. The related TCM terms are explained in the yellow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

*Corresponding author.

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.

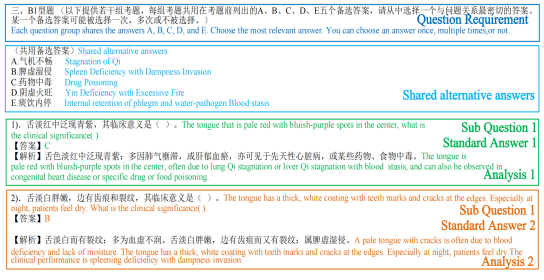

Fig. 3. The example of B1 type of questions. The question requirement is indicated in dark blue text, five options is in light blue text, the first sub question along with the five options, standard answer and analysis is in green text, and the second sub question is in orange text.

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

three question types is relatively uniform.

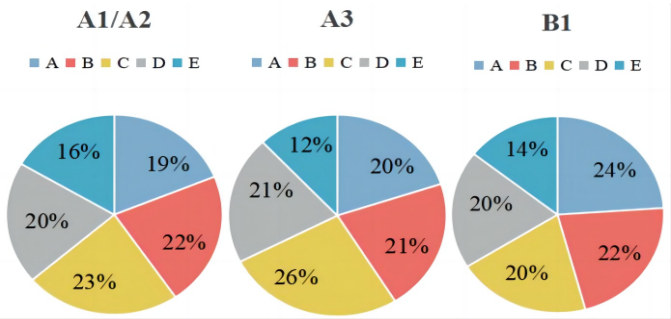

Fig. 4. The distribution of options in TCM-ED among A1/A2 types(left), A3 types(middle), and B1 types(right) of questions.

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<eoe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and<eoa>. (请你一步一步思考并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D，E中选出一个最正确的答案，并写在【答案】和<eoa>之间。)
Example: [Answer]: A <eoa> (例如: 【答案】: A <eoa>)
The format of the complete answer to the question is as follows: (完整的题目回答的格式如下：)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
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A1或者A2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<eoe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and<eoa>. (请你一步一步思考并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D，E中选出一个最正确的答案，并写在【答案】和<eoa>之间。)
Example: [Answer]: A <eoa> (例如: 【答案】: A <eoa>)
The format of the complete answer to the question is as follows: (完整的题目回答的格式如下：)
1)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
2)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
3)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
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A1或者A2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<eoe>. 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<eoa>. (请根据案例所提供的信息进行多轮回答，每一轮都只回答一个问题。请你一步一步思考并将思考过程写在【解析】和<eoe>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中选出一个最正确的答案，并写在【答案】和<eoa>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<eoa>under each question in the example, answer the A3-type questions in the TCMLE. (结合上述示例，特别是根据示例的中每个问题下【解析】与<eoa>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。每组题目有若干道选项相题，共用列出的A、B、C、D、E五个备选答案，请从中选择一个与问题关系最密切的答案。某一个备选答案可能被选择一次，多次或不被选择。)
(CoT and Q&A Constraint)Please think step by step and write your thinking process between [Analysis] and<eoe>. You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and<eoa>. (请你一步一步思考并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D，E中选出一个最正确的答案，并写在【答案】和<eoa>之间。)
Example: [Answer]: A <eoa> (例如: 【答案】: A <eoa>)
The format of the complete answer to the question is as follows: (完整的题目回答的格式如下：)
1)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
2)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
3)
[Analysis]: ... <eoe> (【解析】 ... <eoe>)
[Answer]: ... <eoa> (【答案】 ... <eoa>)
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.