TCM-Ladder: A Benchmark for Multimodal Question Answering on Traditional Chinese Medicine

Jiacheng Xie, Yang Yu, Ziyang Zhang, Shuai Zeng, Jiaxuan He, Ayush Vasireddy, Xiaoting Tang, Congyu Guo, Lening Zhao, Congcong Jing, Guanghui An*, Dong Xu*

NEURAL INFORMATION PROCESSING SYSTEMS

上海中醫藥大學
Shanghai University of Traditional Chinese Medicine
1956

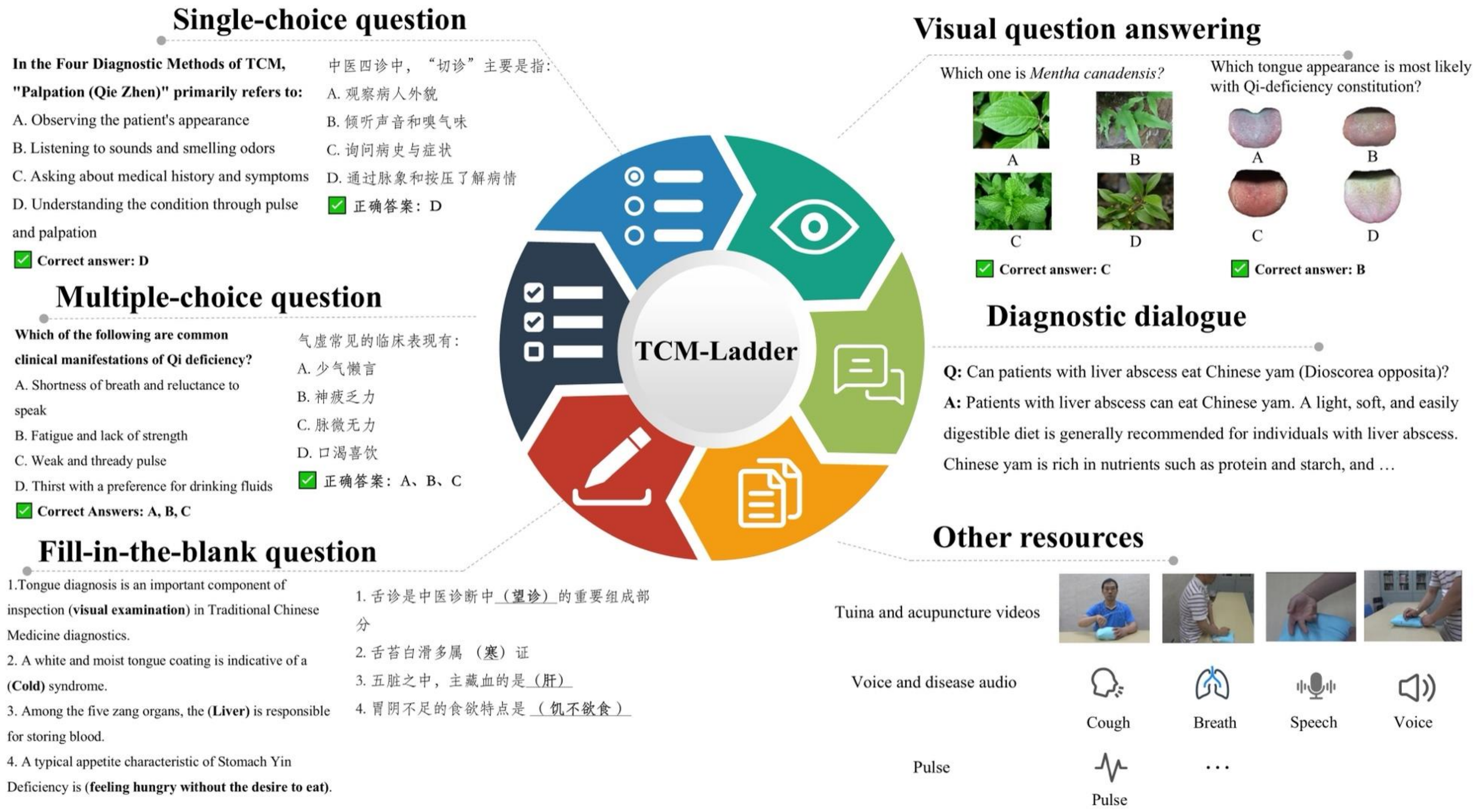
University of Missouri

Overview of TCM-Ladder

TCM-Ladder: a multimodal dataset designed for both training and evaluating TCM-specific and general-domain LLMs.

TCM-Ladder encompasses six task types:

- single-choice questions
- multiple-choice questions
- long-form diagnostic question answering
- fill-in-the-blank tasks
- image-based comprehension task
- additional audio and video resources



Single-choice question

In the Four Diagnostic Methods of TCM, "Palpation (Qie Zhen)" primarily refers to:

A. Observing the patient's appearance

B. Listening to sounds and smelling odors

C. Asking about medical history and symptoms

D. Understanding the condition through pulse and palpation

☑ **Correct answer: D**

中医四诊中，“切诊”主要是指:

A. 观察病人外貌

B. 倾听声音和嗅气味

C. 询问病史与症状

D. 通过脉象和按压了解病情

☑ 正确答案：D

Multiple-choice question

Which of the following are common clinical manifestations of Qi deficiency?

A. Shortness of breath and reluctance to speak

B. Fatigue and lack of strength

C. Weak and thready pulse

D. Thirst with a preference for drinking fluids

☑ **Correct Answers: A, B, C**

气虚常见的临床表现有:

A. 少气懒言

B. 神疲乏力

C. 脉微无力

D. 口渴喜饮

☑ 正确答案：A、B、C

Fill-in-the-blank question

1.Tongue diagnosis is an important component of inspection (**visual examination**) in Traditional Chinese Medicine diagnostics.

2. A white and moist tongue coating is indicative of a (**Cold**) syndrome.

3. Among the five zang organs, the (**Liver**) is responsible for storing blood.

4. A typical appetite characteristic of Stomach Yin Deficiency is (**feeling hungry without the desire to eat**).

1. 舌诊是中医诊断中（望诊）的重要组成部分

2. 舌苔白滑多属 （寒）证

3. 五脏之中，主藏血的是（肝）

4. 胃阴不足的食欲特点是（饥不欲食）

Visual question answering

Which one is *Mentha canadensis?*

☑ **Correct answer: C**

Which tongue appearance is most likely with Qi-deficiency constitution?

☑ **Correct answer: B**

Diagnostic dialogue

Q: Can patients with liver abscess eat Chinese yam (Dioscorea opposita)?

A: Patients with liver abscess can eat Chinese yam. A light, soft, and easily digestible diet is generally recommended for individuals with liver abscess. Chinese yam is rich in nutrients such as protein and starch, and …

Other resources

Tuina and acupuncture videos

Voice and disease audio

Pulse

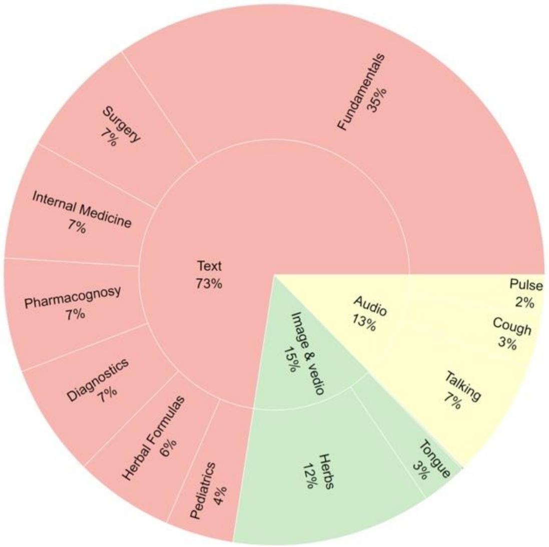
Data Collection and Construction

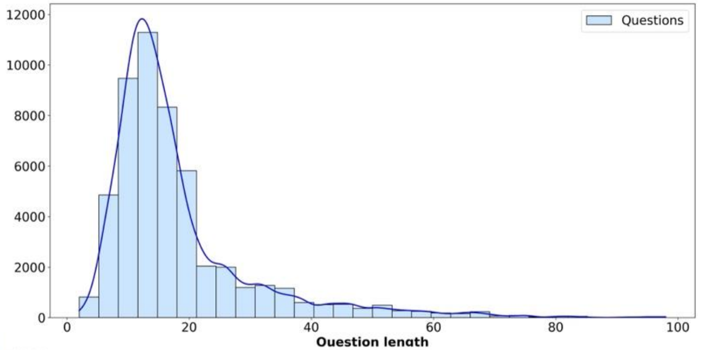
- **Textual QA data**: written by licensed TCM practitioners following a standardized question design protocol and publicly available sources

- **Visual question-answering (VQA):** both manual annotation and automated generation based on existing knowledge bases.
  - **6061 herb images**: from publicly available online resources and photographs we captured at traditional Chinese medicine manufacturing facilities.
  - **1394 tongue images**: collected by a tongue imaging device at Shanghai University of Traditional Chinese Medicine and iTongue software

- **Video data**: recorded by faculty members from the Department of Acupuncture-Moxibustion and Tuina at Shanghai University of Traditional Chinese Medicine

- **Audio data:** publicly available datasets





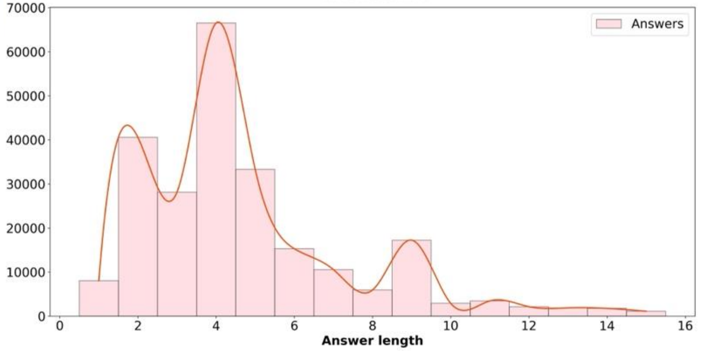


Table 2: Statistics of the collected questions

Statistics	Number
Total questions	52,169
Total answers	238,867
Total subjects	7
Maximum question length	98
Maximum answer length	16
Average question length	18
Average answer length	5
Total images	7,455
Herbs visual questions	6,061
Tongue visual questions	1,394
Total videos	49
Total audios	6,420

Ladder-Score

We introduced **Ladder-Score**, an evaluation metric that integrates TCM-specific terminology and LLM-assisted semantic scoring to assess terminological accuracy and reasoning quality in TCM question answering.

$$\text{Ladder-Score} = \alpha \cdot \text{TermScore} + \beta \cdot \text{SemanticScore}$$

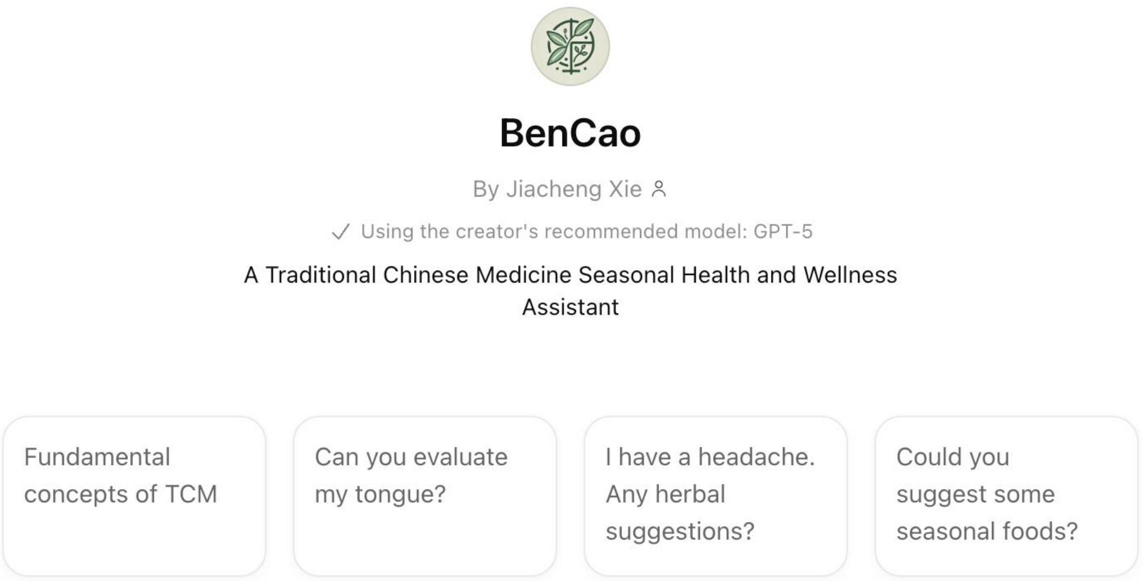
- TermScore, which assesses the accuracy and completeness of TCM terminology usage
- SemanticScore, derived from LLMs to evaluate multiple aspects including logical consistency, semantic accuracy, comprehensiveness of knowledge, and fluency of expression
- α = 0.4, β = 0.6 ,which can be adjusted based on practical needs.

Model Training

We trained two models using the TCM-Ladder dataset:

- **BenCao** an online model fine-tuned from ChatGPT. BenCao was trained on knowledge extracted from over 700 classical Chinese medicine books, none of which contained any question-answer pairs.

- **Ladder-base**, which is built upon the pretrained Qwen2.5-7B-Instruct model and enhanced with Group Relative Policy Optimization (GRPO) to improve its reasoning capabilities.



BenCao

By Jiacheng Xie

✓ Using the creator's recommended model: GPT-5

A Traditional Chinese Medicine Seasonal Health and Wellness Assistant

Fundamental concepts of TCM
Can you evaluate my tongue?
I have a headache. Any herbal suggestions?
Could you suggest some seasonal foods?

Benchmark Results

Text-Based Single and Multiple-Choice Question Answering

- We evaluated **nine state-of-the-art general-domain LLMs** and **five TCM-specific models**

- **Ladder-base** consistently outperforms other models across all subject areas, achieving the highest overall accuracy.

- Our model, **BenCao**, also demonstrates robust performance, particularly in Diagnostics and Internal Medicine.

- Gemini 2.5 Pro, Deepseek, and Qwen3 show relatively stable accuracy across domains, with scores ranging from 0.65 to 0.75, though they still fall short compared to domain-specific models.

- Claude 3, GPT-4o mini, and BenTsao underperform.

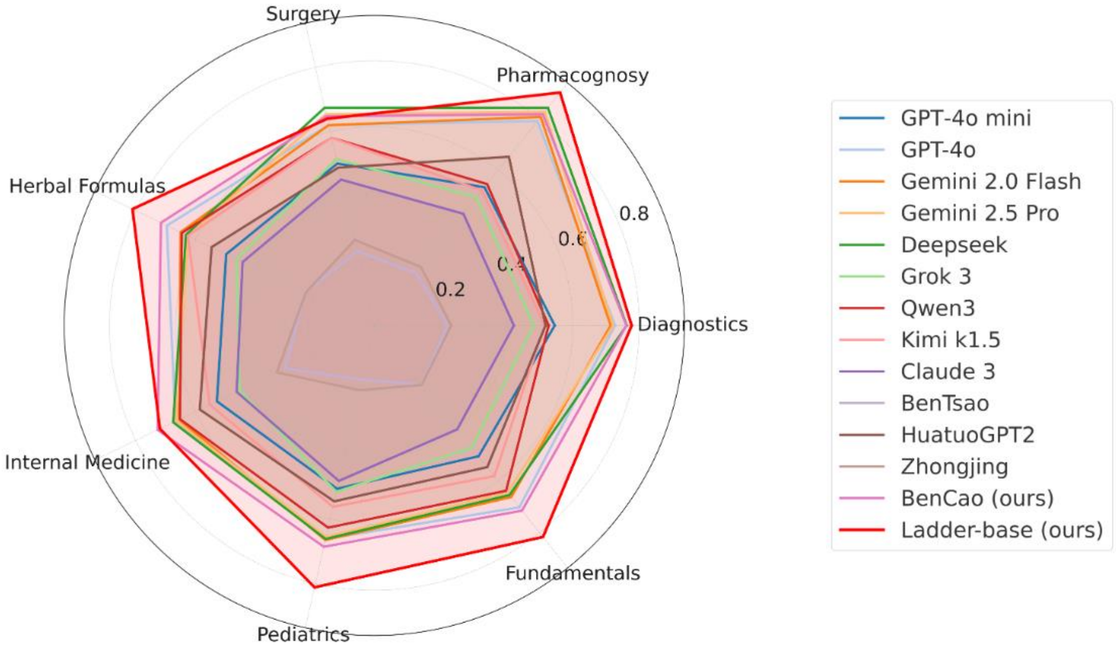


Figure 3: Performance of general-domain and TCM-specific language models on single and multiple-choice question answering tasks.

Benchmark Results

Visual Question Answering

- To further assess the models' capability in visual understanding tasks within TCM, we evaluated ten LLMs on two image-based benchmarks: herbs classification and tongue image diagnosis.

- **BenCao** achieves the highest accuracy in both tasks.

- General-domain LLMs such as Gemini 2.5 Pro, Gemini 2.0 Flash, and Qwen3 exhibit moderate performance.

- In contrast, models like GPT-4o, Claude 3, Kimi k1.5, and Grok 3 demonstrate limited performance, particularly in the tongue classification task.

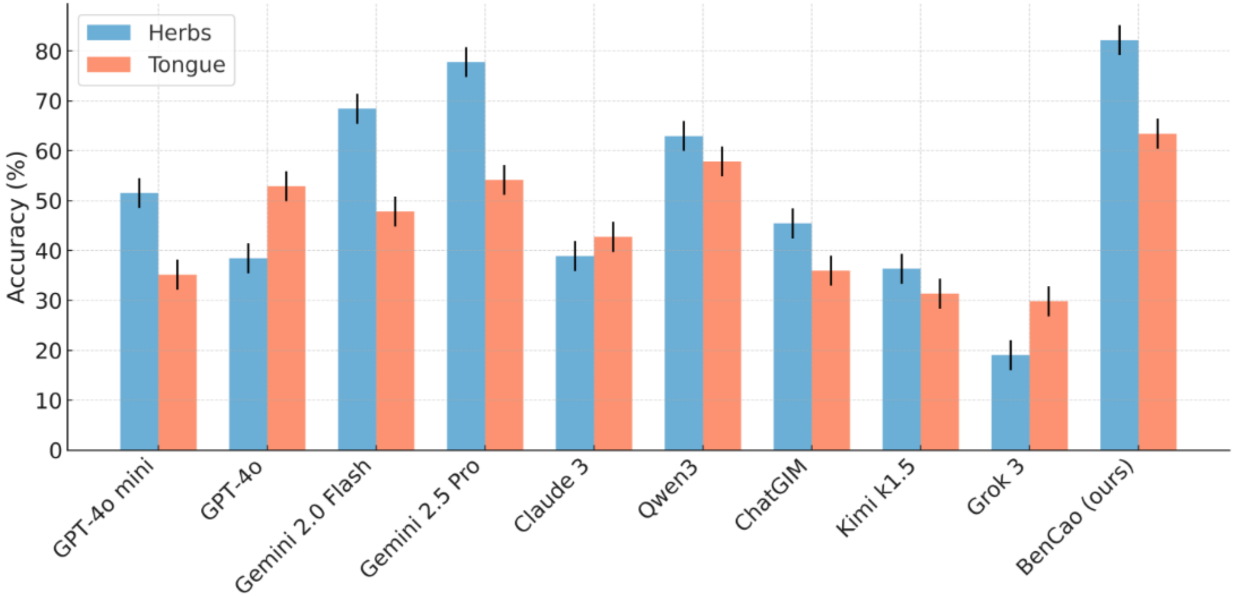


Figure 4: The performance of large language models on questions regarding Chinese herbal medicine and tongue image classification.

Benchmark Results

Diagnostic Dialogue and Fill-in-the-Blank Questions

- In the **diagnostic dialogue task**, our model Ladder-base achieved the highest scores in BLEU-4 and ROUGE-L and also maintaining a strong Ladder-Score.
- Qwen3 achieved the best Ladder-Score and the highest METEOR.
- BenCao achieved the best BERTScore.
- In the **fill-in-the-blank task**, BenCao significantly outperformed all other models, achieving the highest exact match accuracy

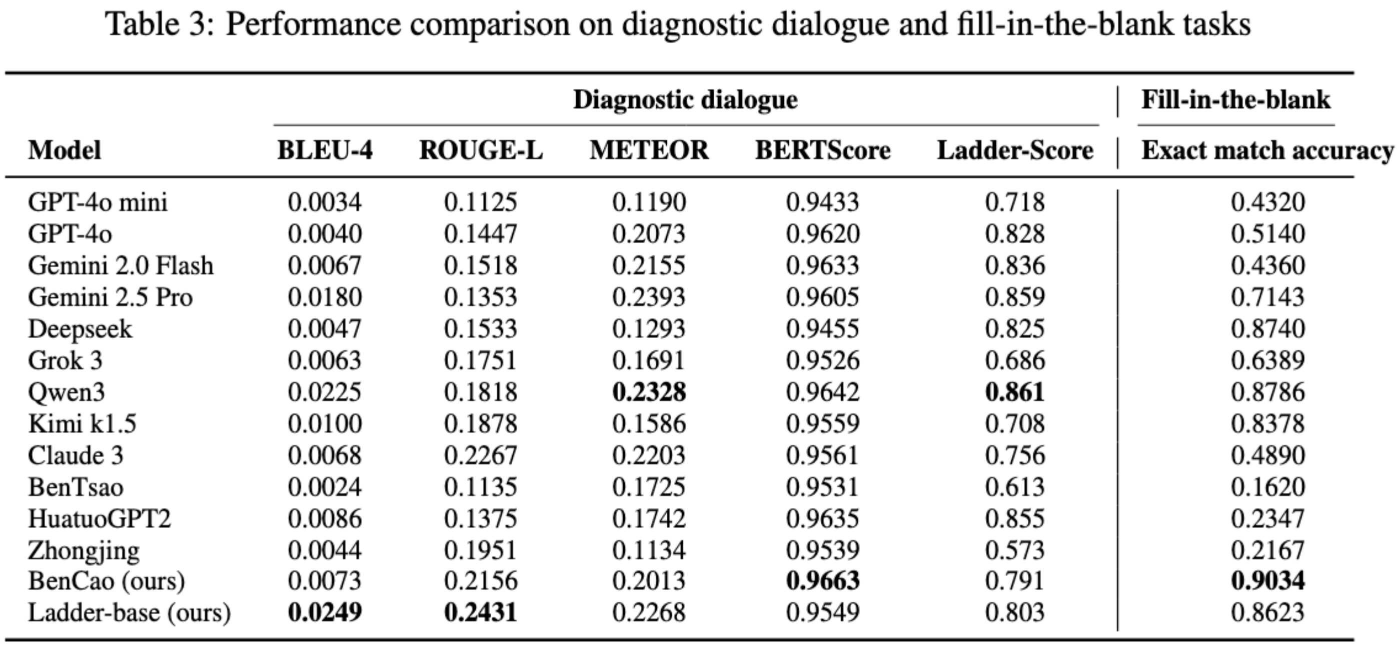
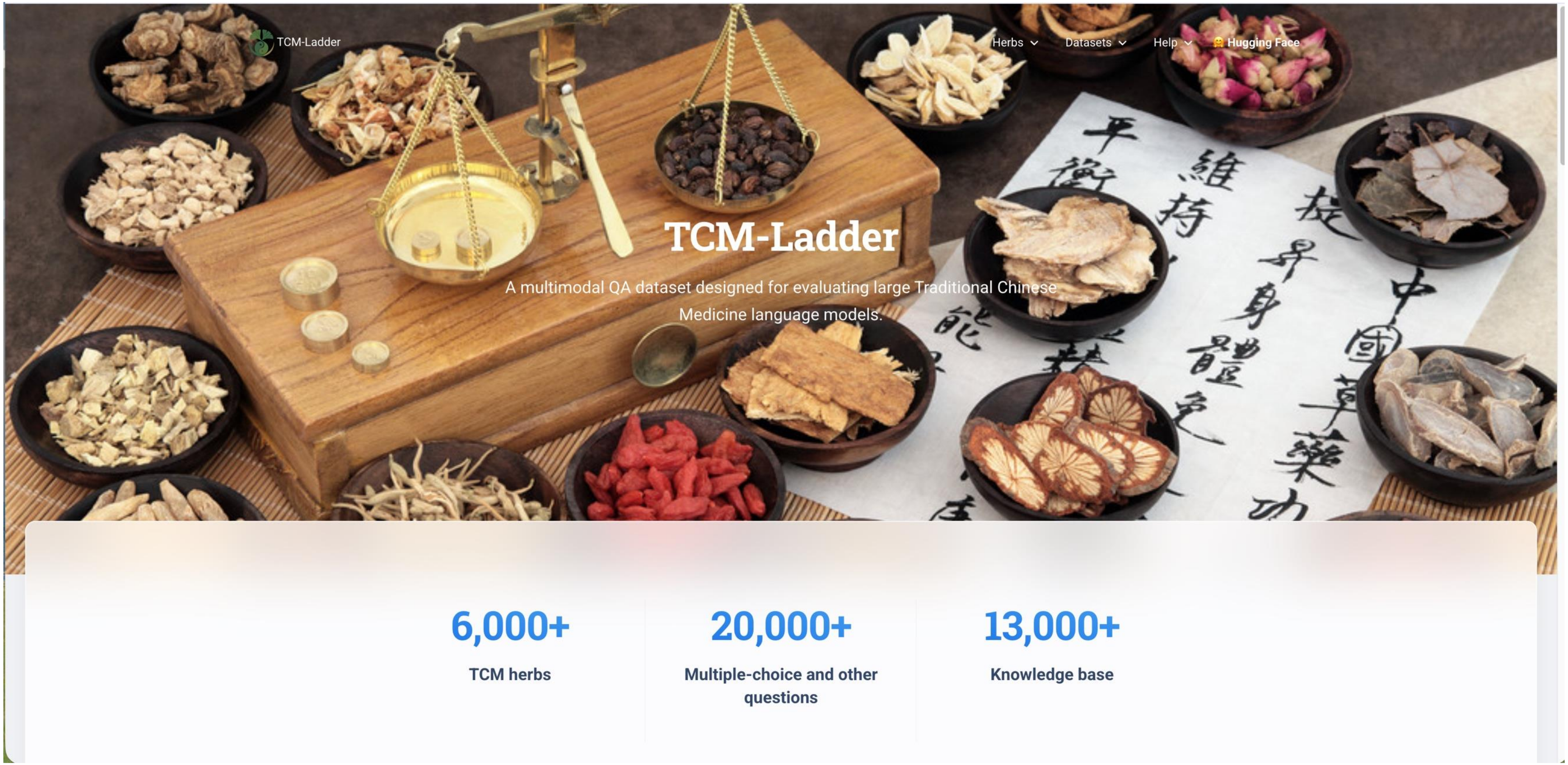


Table 3: Performance comparison on diagnostic dialogue and fill-in-the-blank tasks

	Diagnostic dialogue					**Fill-in-the-blank**
Model	**BLEU-4**	**ROUGE-L**	**METEOR**	**BERTScore**	**Ladder-Score**	**Exact match accuracy**
GPT-4o mini	0.0034	0.1125	0.1190	0.9433	0.718	0.4320
GPT-4o	0.0040	0.1447	0.2073	0.9620	0.828	0.5140
Gemini 2.0 Flash	0.0067	0.1518	0.2155	0.9633	0.836	0.4360
Gemini 2.5 Pro	0.0180	0.1353	0.2393	0.9605	0.859	0.7143
Deepseek	0.0047	0.1533	0.1293	0.9455	0.825	0.8740
Grok 3	0.0063	0.1751	0.1691	0.9526	0.686	0.6389
Qwen3	0.0225	0.1818	**0.2328**	0.9642	**0.861**	0.8786
Kimi k1.5	0.0100	0.1878	0.1586	0.9559	0.708	0.8378
Claude 3	0.0068	0.2267	0.2203	0.9561	0.756	0.4890
BenTsao	0.0024	0.1135	0.1725	0.9531	0.613	0.1620
HuatuoGPT2	0.0086	0.1375	0.1742	0.9635	0.855	0.2347
Zhongjing	0.0044	0.1951	0.1134	0.9539	0.573	0.2167
BenCao (ours)	0.0073	0.2156	0.2013	**0.9663**	0.791	**0.9034**
Ladder-base (ours)	**0.0249**	**0.2431**	0.2268	0.9549	0.803	0.8623

Application Website

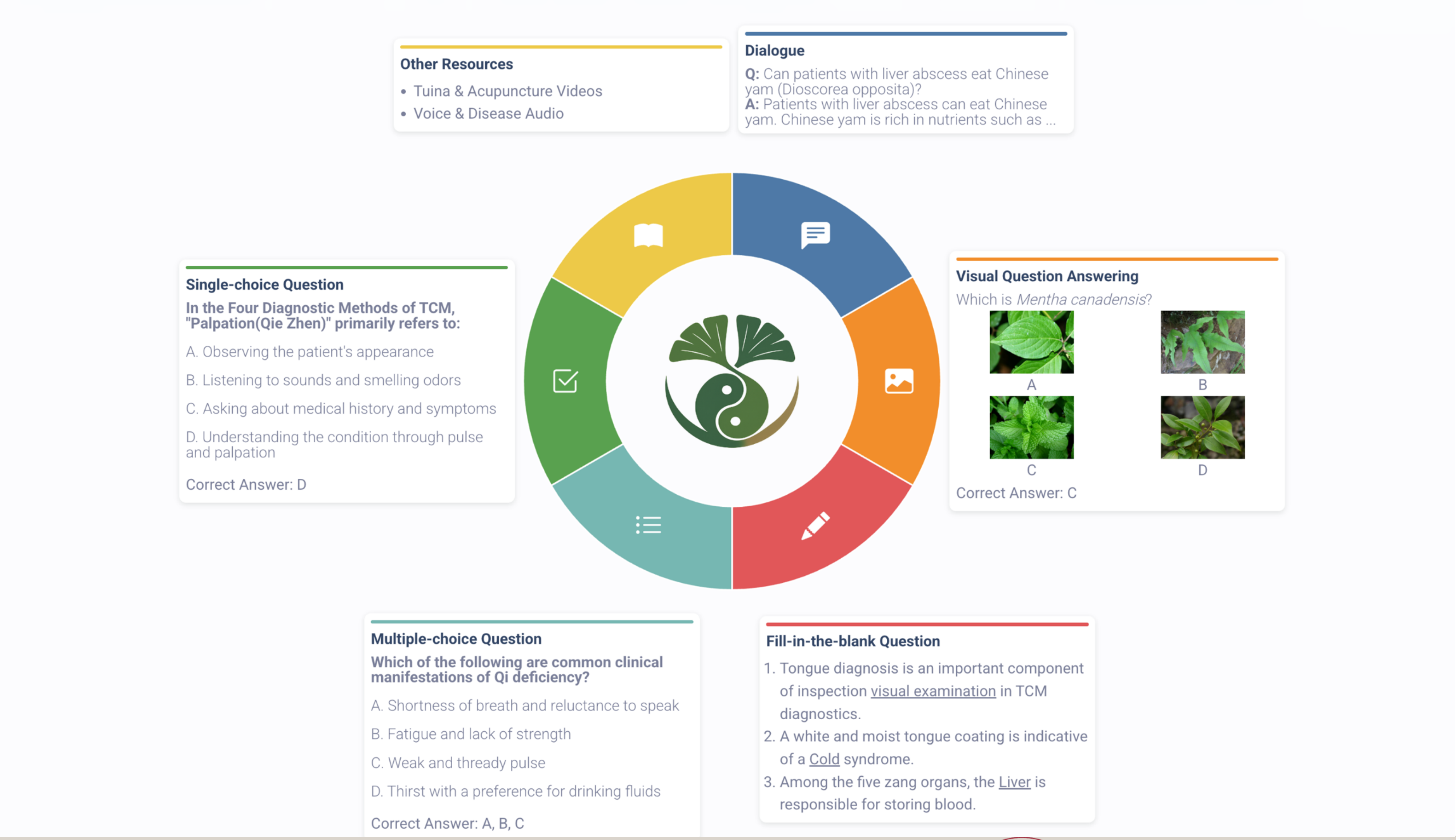


TCM-Ladder

A multimodal QA dataset designed for evaluating large Traditional Chinese Medicine language models.

6,000+	20,000+	13,000+
TCM herbs	Multiple-choice and other questions	Knowledge base

Application Website



Other Resources

- Tuina & Acupuncture Videos
- Voice & Disease Audio

Dialogue

Q: Can patients with liver abscess eat Chinese yam (Dioscorea opposita)?
A: Patients with liver abscess can eat Chinese yam. Chinese yam is rich in nutrients such as ...

Single-choice Question

In the Four Diagnostic Methods of TCM, "Palpation(Qie Zhen)" primarily refers to:

A. Observing the patient's appearance

B. Listening to sounds and smelling odors

C. Asking about medical history and symptoms

D. Understanding the condition through pulse and palpation

Correct Answer: D

Visual Question Answering

Which is *Mentha canadensis*?

Correct Answer: C

Multiple-choice Question

Which of the following are common clinical manifestations of Qi deficiency?

A. Shortness of breath and reluctance to speak

B. Fatigue and lack of strength

C. Weak and thready pulse

D. Thirst with a preference for drinking fluids

Correct Answer: A, B, C

Fill-in-the-blank Question

1. Tongue diagnosis is an important component of inspection visual examination in TCM diagnostics.
2. A white and moist tongue coating is indicative of a Cold syndrome.
3. Among the five zang organs, the Liver is responsible for storing blood.