A Availability

- You can access TCM-Ladder website at `http://tcmladder.com`
- The GitHub repository with evaluation code and prompts is available at `https://github.com/orangeshushu/TCM-Ladder`
- Data can be downloaded from `https://huggingface.co/datasets/timzzyus/TCM-Ladder`

B Data resources

As shown in Table 1 and Table 2, we incorporated several publicly available resources from existing publications to extend and supplement the voice and pulse diagnosis datasets. For each dataset, we provide the corresponding original publication reference and licensing documentation. In addition, we partially incorporated several publicly available long-form dialogue datasets to supplement our corpus. The sources and licensing statements of these datasets are provided in Table 3.

Table 1: Summary of publicly available voice and disease audio datasets

Dataset	Modality	Language	Task Type	Number of Samples	Year	License
COUGHVID[1]	Cough audio	English	COVID-19 detection	20,000	2020	CC-BY 4.0
Coswara[2]	Cough, breath, speech	English	COVID-19 detection	5,000	2022	CC-BY 4.0
UK COVID-19 Vocal Audio Dataset[3]	Cough, breath, speech	English	COVID-19 detection	70,000	2023	OGL v3.0
Respiratory Sound Database[4]	Lung auscultation sounds	English	Respiratory disease classification	920	2017	CC-BY 4.0
smarty4covid[5]	Cough, breath, voice	English	COVID-19 detection	4,600	2023	CC-BY 4.0
Bridge2AI-Voice[6]	Voice recordings	English	Voice biomarker research	Not specified	2025	Apache-2.0
VOICED[7]	Voice recordings	English	Pathological voice analysis	208	2018	ODC-BY 1.0
Perceptual Voice Qualities Dataset[8]	Voice recordings	English	Perpetual voice quality	360+	2020	CC-BY 4.0
COVID-19 Voice Dataset[9]	Voice recordings	English	COVID-19 detection	Not specified	2023	CC-BY 4.0
ALS IAC Speech Corpus[10]	Speech	English	ALS	Not specified	2024	CC-BY 4.0
PMC COVID-19 Voice Dataset[11]	Voice recordings	English	COVID-19 detection	Not specified	2022	OGL v3.0

Table 2: Summary of publicly available pulse datasets

Dataset	Modality	Language	Task Type	Number of Samples	Year	License
PulseDB[12]	ECG, PPG, ABP waveforms	English	Cuff-less blood pressure estimation	5,245,454	2023	ODbL
MIMIC-BP[13]	ECG, PPG, ABP waveforms	English	Blood pressure estimation	12,000	2024	ODC-By 1.0
Pulse-ECG[14]	ECG images	English	ECG interpretation	1,160,000	2023	Apache-2.0
MTHS Dataset[15]	Video-PPG, ECG signals	English	Heart rate and SpO2 estimation	65	2023	CC BY-NC-ND 4.0
Welltory Dataset[16]	Video-PPG, ECG signals	English	Heart rate variability analysis	21	2023	CC BY-NC-ND 4.0
BUT-PPG Dataset[17]	Video-PPG signals	English	Heart rate estimation	65	2023	CC-BY 4.0

Table 3: Summary of available long-form TCM dialogue dataset

Dataset	Modality	Language	Task Type	Number of Samples	Year	License
Huatuo-26M[18]	Text	Chinese	QA, Dialogue	26M+	2023	CC-BY 4.0
TCMD[19]	Text	Chinese	Syndrome-Finding Mapping	100,000+	2024	CC-BY 4.0
CMD[20]	Text	Chinese	Medical Dialogue	25,000+ dialogues	2020	MIT

C Training details

For *Bencao*: *Bencao* is a model fine-tuned on the basis of a large language model, GPT (Generative Pre-trained Transformer)[21]. The fine-tuning process leveraged a curated corpus of over 700 TCM classical books as its knowledge base. Reinforcement Learning from Human Feedback (RLHF)[22] was employed to iteratively improve the model's output quality, with human annotators evaluating the helpfulness and accuracy of its responses.

The model was initially guided by role-defining prompts. For example, one representative instruction states: "*You are an experienced TCM expert. Please respond to users' questions based on your expertise and the content in your knowledge base. Try to combine traditional TCM terminology*

with modern language to ensure both professionalism and clarity. Do not provide any diagnostic conclusions, and explicitly mention the limitations of your responses."

To refine the model's output, we conducted more than 200 rounds of multi-turn instructional fine-tuning. This iterative process aimed to ensure that the model's responses were as professional, objective, and informative as possible.

It is important to note that the training corpus consisted primarily of authoritative classical TCM literature and educational textbooks. Some associated books also included tongue images and herbal illustrations. However, no question-answer pairs from the TCM-Ladder benchmark were used in any stage of the training or fine-tuning process.

For *Ladder-base*: The Ladder-base model is fine-tuned on the *Qwen2.5-7B* model using Group Relative Policy Optimization(GRPO)[23], with the train set in TCM-Ladder. A consistent system prompt is used for all questions: "*You are a helpful AI Assistant that provides well-reasoned and detailed responses. You first think about the reasoning process as an internal monologue and then provide the user with the answer. The response cannot be more than 400 words. Each question has only one answer, A-E. Respond in the following format:* <think>\n...\n</think> \n<answer>\n...Answer:A-E...\n</answer>. *Do not respond with another tag. Do not repeat the response.*". The final response is parsed and evaluated using a rule-based reward system that assigns one point each for a correct answer, proper format, and correct tags. These rewards are weighted in a ratio of 5:1:1 to encourage the model to prioritize answering questions correctly.

For a question–answer pair (q, a), the policy model $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_i\}_{i=1}^{G}$. The advantage of the i-th response is calculated by normalizing the group-level reward $\{R_i\}_{i=1}^{G}$

$$\hat{A}_{l,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})} \tag{1}$$

GRPO first computes the mean loss across each generated sequence and then averages these losses over all sampled sequences. It employs a clipped objective function combined with an explicitly applied KL divergence penalty term between policy model and reference model:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Big[q \sim P(Q),\ \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid q)\Big] \\ \cdot \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t},\right.\right. \\ \left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})},\ 1-\epsilon,\ 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right\}\end{aligned} \tag{2}$$

D Evaluation environment

For general domain large language models such as *GPT-4*, *GPT-4o mini*, *Gemini 2.0 Flash*, *Gemini 2.5 Pro*, and *Deepseek*, we conducted evaluations using their respective APIs. The corresponding code is available in our GitHub repository. For TCM-specific models such as *HuatuoGPT*[24], *Zhongjing*[25], and *Bentsao*[26], we performed local evaluations, and the testing code for these models is also provided in the GitHub repository to facilitate reproducibility.

All experiments were conducted on the Hellbender[27] computing cluster at the University of Missouri. Detailed configurations of the computational resources are provided in Table 4.

E Safeguards

All released images underwent thorough manual review, and tongue images were subjected to strict de-identification procedures to ensure that no personally identifiable information is included. Use of our dataset requires agreement to the terms of use provided.

When using our released models, users are provided with appropriate risk disclaimers. For example: the responses generated by this model are for informational purposes only and should not be con-

Table 4: Detailed configuration of computational resources

Model	Dell R7 40xa
Nodes	17
Cores/node	64
System Memory	238 G
GPU	A100
GPU Memory	80 GB
Numbers of GPUs	4
Local Scratch	1.6 TB

sidered a substitute for professional medical diagnosis. If you experience any discomfort or health concerns, please seek medical attention promptly.

F Declaration of LLM usage

We employed LLMs as a core component of our evaluation framework. Specifically, both general-domain and TCM-specific LLMs were systematically evaluated on *TCM-Ladder*, a multimodal benchmark designed for TCM. For multiple-choice questions, we used standardized zero-shot prompts, for example: "*Please answer the following multiple-choice question and clearly indicate the letter (A, B, C, D, or E) of the option you believe is correct.*" We provide implementation code for each evaluated LLM to ensure transparency and reproducibility (see GitHub links in **Appendix A**). Furthermore, LLMs are integral to our proposed textual evaluation metric, *Ladder-Score*, where LLMs assess model-generated answers based on well-defined rubrics. These rubrics span multiple dimensions, including logical consistency, semantic accuracy, knowledge coverage, and fluency of expression. Detailed evaluation protocols and rubric definitions are described in **Appendix H**.

G Herb and tongue image data collection

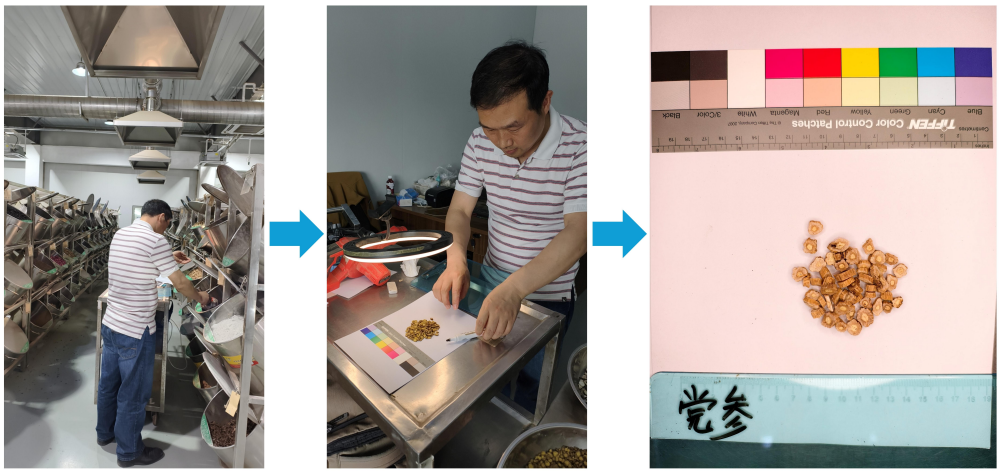
The images of Chinese medicinal herbs were sourced from two primary channels: a portion was collected from publicly accessible online resources, while the remainder was obtained through on-site photography conducted at *Shanghai Kangqiao Pharmaceutical Co., Ltd*[28]. As illustrated in Figure 1, we first identified each herb within the pharmaceutical facility, followed by standardized image capture under controlled lighting conditions. Additionally, a measurement scale and a standard color chart were placed alongside each herb specimen to facilitate subsequent analytical processing and color correction.

Tongue image acquisition was conducted through two primary methods. As shown in Figure 2, a portion of the images was collected using a specialized tongue imaging device[29], which provides a stable lighting environment to ensure image consistency. Another portion was obtained via our mobile application, *iTongue*[30] (as shown in Figure 3), which enables analysis and prediction of TCM body constitution. To maximize the protection of participants' privacy, only the tongue surface region of the images has been released publicly.

We use the existing image labels as ground truth answers and select three incorrect options as distractors. Based on specific question templates such as "*Which of the following images shows the herb [Herb Name]?*", we generate corresponding visual questions. The code for generating these visual questions is publicly available on GitHub.

H Ladder-Score: A hybrid evaluation metric for long-form TCM question Answering

To better evaluate the quality of long-form answers generated by TCM-specific LLMs in TCM settings, we introduce *Ladder-Score*, a hybrid evaluation metric that combines terminology matching with rubric-based semantic assessment.



(a) Collection of Chinese herbs (b) Image acquisition (c) Sample image (with scale and color card)

Figure 1: The image acquisition process of Chinese medicinal herbs. (a) Collection of physical herb specimens from a pharmaceutical factory. (b) Image capture under controlled lighting conditions. Photographs of the herbs were taken to ensure consistency and quality. (c) A sample herb image, including a scale bar and a color calibration card, which facilitates subsequent processing and color correction.

(a) Collected tongue image samples (b) Acquisition process using a tongue imaging device

Figure 2: Tongue image acquisition process using a tongue imaging device. (a) Sample tongue images collected from participants. To protect subject privacy, the eye regions have been masked. (b) Demonstration of the tongue image capture procedure using the imaging device.

H.1 Terminology-based score (*TermScore*)

We construct a domain-specific term list using *TCMBank*[31], international TCM terminology standards[32], and manually curated clinical terms (e.g., syndromes, symptoms, treatment principles, herbal formulas). Given a reference answer R and a generated candidate answer C, we extract their term sets T_R and T_C via Jieba-based tokenization and lexicon matching.

We compute:

- **Term Coverage**: the proportion of reference terms present in the candidate:

$$TC = \frac{|T_C \cap T_R|}{|T_R|} \tag{3}$$

- **Term Consistency**: the semantic similarity between term sets using sentence-level BERT embeddings and cosine similarity.

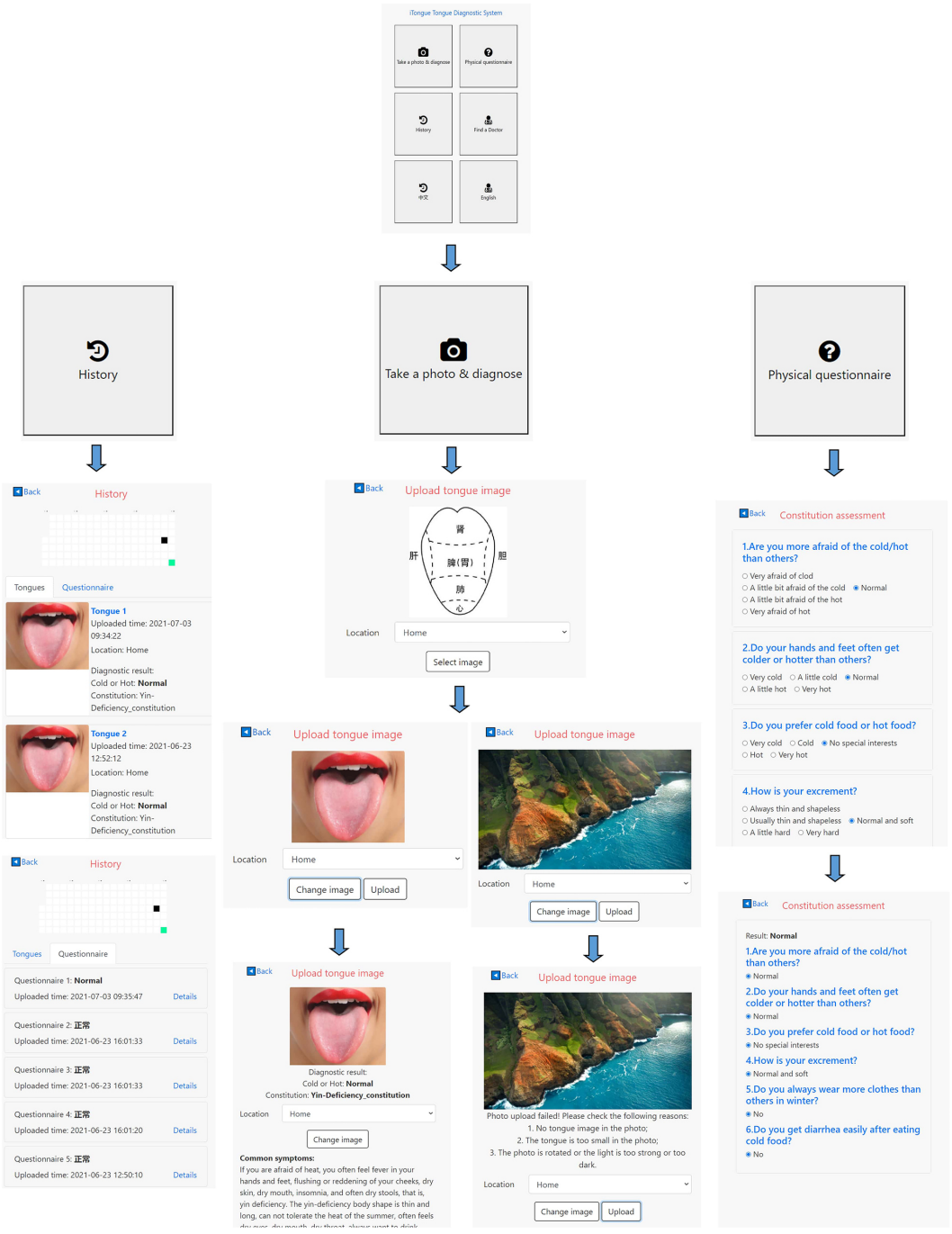


Figure 3: User interaction with *iTongue* app. Users can upload their own tongue images to receive TCM-based assessments, including cold–heat syndrome differentiation and body constitution classification. If a non-tongue image is uploaded, the application prompts the user to re-upload a valid tongue image.

The overall *TermScore* is:

$$\textit{TermScore} = \lambda \cdot TC + (1 - \lambda) \cdot TS \quad (\lambda = 0.5) \tag{4}$$

H.2 Semantic rubric score (*SemanticScore*)

To assess semantic quality, we prompt an external LLM (e.g., GPT-4 or Claude) using a rubric covering four dimensions as shown in Table 5.

Table 5: The four dimensions of the LLM scoring rubric.

Dimension	***Description***
Diagnostic Accuracy	Whether a correct syndrome or pathology is inferred
Treatment Appropriateness	Whether the recommendation aligns with clinical reasoning
Logical Coherence	Whether the answer is structured and internally consistent
Medical Language Quality	Clarity, fluency, and use of professional language

Each is rated on a 0–5 scale. The *SemanticScore* is the normalized average:

$$\textit{SemanticScore} = \frac{1}{20} \sum_{i=1}^{4} s_i \tag{5}$$

We apply blind prompting to prevent model self-bias, and optionally average across multiple LLM reviewers.

H.3 Final metric: Ladder-Score

The final score is a weighted sum of both components:

$$\text{Ladder-Score} = \alpha \cdot \text{TermScore} + \beta \cdot \text{SemanticScore} \tag{6}$$

This formulation ensures both domain fidelity and semantic quality are captured in evaluating long-form generative answers under TCM tasks.

H.4 LLM-Based evaluation prompt template

To elicit structured ratings from LLMs, we use the following prompt template for each generated answer:

"You are a TCM expert evaluating a model-generated answer. Please score the answer according to the following rubric, from 0 (very poor) to 5 (excellent):

1. Diagnostic Accuracy: Is the TCM syndrome or pathology properly inferred?

2. Treatment Appropriateness: Are the therapeutic suggestions clinically relevant and reasonable?

3. Logical Coherence: Is the explanation consistent and well-structured?

4. Medical Language Quality: Is the response clearly written with accurate terminology?

Here is the candidate answer:

[candidate answer here]

Return your scores in the following format:

Diagnostic Accuracy: [0–5]

Treatment Appropriateness: [0–5]

Logical Coherence: [0–5]

Medical Language Quality: [0–5]"

H.5 Terminology extraction strategy

To compute the *TermScore*, we extract domain-specific terms from both reference and candidate answers using the following procedure:

Lexicon construction:

We build a term dictionary combining:

- The *TCMBank*[31] herb/syndrome/formula corpus
- Standardized clinical terminology lists
- Manual augmentation with high-frequency terms from training corpora

Tokenization and filtering:

- Use *jieba*[33] for Chinese segmentation.
- Retain terms satisfying:
  - Length ≥ 2
  - Occurrence in the constructed lexicon or inclusion of suffixes.

Semantic similarity (for Term Consistency):

- Compute sentence/term embeddings using pretrained Chinese BERT[34].
- Measure cosine similarity between aggregated term vectors from references and candidates.
- Use thresholded or averaged score as term semantic match score.

H.6 Case study

To assess the effectiveness of the proposed *Ladder-Score*, we compared its performance against commonly used automatic evaluation metrics including BLEU-4[35], ROUGE-L[36], METEOR[37], and BERTScore[38]. While these traditional metrics primarily rely on surface-level lexical overlap or embedding similarity, they fail to capture the nuanced reasoning, terminology accuracy, and domain alignment that are critical in TCM long-form question answering.

For instance, for the question (English translation):"A patient presents with a red tongue, yellow coating, bitter taste in the mouth, restlessness, and dark yellow urine. What is your diagnosis and treatment plan according to TCM?"

Reference answer (Expert-written) is:"Based on the symptoms of red tongue with yellow coating, bitter taste, and restlessness, the pattern can be identified as Liver-Gallbladder Damp-Heat. The treatment principle is to clear heat and drain dampness while soothing the Liver. The classical formula Long Dan Xie Gan Tang is recommended, with herbs like Gentiana, Scutellaria, and Gardenia. Patients should avoid greasy, spicy food to prevent worsening of symptoms."

A high-quality candidate answer A generated by a large language model is:"The clinical presentation suggests Liver and Gallbladder damp-heat syndrome, often due to emotional stagnation and internal dampness. Treatment should focus on clearing heat, resolving dampness, and soothing the Liver. Long Dan Xie Gan Tang is appropriate, including herbs like Gentiana, Alisma, Plantago Seed, and Scutellaria. In cases with headaches or pronounced bitterness in the mouth, Bupleurum and Mint can be added."

Another candidate answer B, generated by a different model, exhibits high lexical overlap but lacks substantive content:"Red tongue, yellow coating, and bitter taste indicate Liver-Gallbladder damp-heat. You can use Long Dan Xie Gan Tang for treatment. This includes herbs like Gentiana, Gardenia, and Scutellaria. The patient should avoid spicy food."

We computed the evaluation scores of the above responses using different metrics, as shown in Table 6. In this diagnostic task, the two model-generated responses may share a significant amount of surface-level terminology (e.g., "Long Dan Xie Gan Tang," "bitter taste," "avoid spicy food"). However, one response merely paraphrases the reference answer, while the other provides a precise pattern differentiation, pharmacological rationale, and therapeutic strategy. Despite the substantial difference

in clinical quality, metrics such as BLEU and ROUGE may yield similar scores due to their reliance on lexical overlap. In contrast, *Ladder-Score* demonstrates a significant advantage. By integrating terminology matching with a rubric-based semantic evaluation mechanism, it can accurately capture the following aspects: (1) the accuracy of pattern differentiation; (2) the appropriateness of formula and herb recommendations; (3) the normative use of TCM-specific terminology; (4) the presence of templated or generic content; and (5) the ability to robustly assess complex clinical responses in long-form text.

This dual evaluation mechanism enables *Ladder-Score* to more reliably reflect the practical utility of model outputs in professional clinical settings. It is particularly well suited for specialized tasks that demand not only medical accuracy but also reasoning capability and contextual alignment, capabilities that are often lacking in conventional evaluation metrics.

Table 6: Performance of the two candidate responses under different evaluation metrics.

Metric	Candidate A	Candidate B	Difference	Can it distinguish quality?	Remarks
BLEU-4	0.45	0.43	+0.02	No	Rewards word overlap; ignores reasoning depth
ROUGE-L	0.52	0.51	+0.01	No	Captures sequence overlap; surface-level only
METEOR	0.44	0.45	–0.01	No	Slightly penalizes varied but correct expressions
BERTScore	0.87	0.85	+0.02	No	Sensitive to semantics, but lacks clinical awareness
Ladder-Score	**0.91**	**0.75**	**+0.16**	**Yes**	**Correctly rewards richer clinical reasoning and term use**

I Question design protocol and selection

When manually composing questions, TCM practitioners adhered to the criteria outlined in Table 7. The question stems were constructed based on content from official TCM textbooks, the *National Medical Licensing Examination syllabus*, or relevant online knowledge bases. During the question development process, care was taken to ensure that the stems were concise, unambiguous, and clearly stated. The questions were designed to cover multiple cognitive levels, including factual recall, conceptual understanding, and clinical application, while maintaining a diverse range of difficulty levels. Upon completion, the questions were further reviewed and filtered according to the screening criteria specified in Table 8 to eliminate redundancy and ensure quality.

Table 7: Guidelines and standards for constructing questions.

Category	Key points
Source and reference	- Base questions on authoritative sources such as official TCM textbooks, national curricula, and licensing examination syllabi. - Ensure coverage of fundamental TCM subjects, including fundamentals, diagnostics, herbal formulas, and pediatrics. - Utilize reputable databases for reference, such as *TCM Bank*[31], *ETCM*[39], *HERB 2.0*[40], etc.
Question stem design	- Ensure clarity, precision, and absence of ambiguity in wording. - Focus each question on a single knowledge point, avoiding irrelevant information. - Employ diverse question types to avoid formulaic stems.
Option construction	- Ensure only one correct answer for single-choice items; provide clearly defined correct combinations for multiple-choice items. - Include plausible distractors that test the depth of understanding. - Avoid leading cues such as key terms from the stem repeated in the correct answer.
Difficulty level	- Include questions across a range of difficulties. - Encourage the use of comprehension and application-based questions.
Common pitfalls to avoid	- Avoid content that is beyond the TCM curriculum scope or lacks clear relevance. - Prioritize clinical reasoning and critical analysis. - Ensure grammatical consistency between stem and options. - Avoid ambiguous, logically flawed, or duplicate questions.
Review and validation	- All items should undergo peer review by qualified TCM educators or clinicians. - Employ automated tools to detect duplications and evaluate ambiguity.

In addition to manual duplication checks, we employed three duplicate detection methods: string edit distance[41], TF-IDF[42, 43] with cosine similarity, and BERT [44, 45] based semantic encoding, to achieve multi level duplicate detection from surface to semantic levels. We considered two candidate questions to be duplicates if their similarity scores exceeded a threshold of 0.9 in any of the three methods mentioned above.

Table 8: Criteria for the question selection.

Dimension	Detailed explanation
Duplicate question	If two or more questions within the same category assess the same knowledge point with only slight variations in phrasing, retain only the version with clearer and more complete wording.
Ambiguous question	Any question with ambiguous wording, unclear intent, poor grammar, or syntactic confusion should be considered invalid and removed.
Answer uncertainty handling	If a single-choice question has multiple plausible answers or lacks clear differentiation between correct and distractor options, it should be removed. Similarly, multi-choice questions without a clear key or option logic should be discarded.
Option structure issues	Questions should be removed or revised if options are repetitive, nested within each other, or display length bias (e.g., one option is significantly longer than others).
Format and convention check	Questions missing prompts, unnumbered options, or nonstandard language/punctuation should be revised. If unfixable, they should be eliminated.
Outdated question context	Questions that conflict with the latest clinical guidelines or research, such as outdated treatment plans, should be revised or removed.

J Limitations and societal impact

TCM-Ladder serves as an evaluation benchmark for both general domain LLMs and TCM-specific models, offering a multimodal dataset to support future training and evaluation efforts in the field of TCM. However, several limitations remain and warrant further improvement and refinement in subsequent work.

1. **Limitations in multimodal evaluation**. The current models' multimodal capabilities have been evaluated only on visual tasks involving herb images and tongue diagnosis images. Other data modalities that are also crucial in TCM, such as pulse diagnosis signals, audio (e.g., patient voice), and video, have not yet been incorporated into the evaluation framework in this study.

2. **Semantic deviations in cross-lingual translation**. During the construction of the multilingual version of the dataset, some Chinese questions may have incurred semantic deviations or ambiguities during the translation into English. Given the context-dependent and culturally embedded nature of TCM terminology, some expressions do not have direct equivalents in English, which can impair the model's comprehension and reasoning capabilities. Such translation errors are particularly pronounced in complex dialectical reasoning tasks and may reduce the reliability of model evaluation. Future iterations of the dataset should involve closer collaboration between linguistic experts and TCM practitioners to ensure higher consistency and fidelity in cross-lingual representation.

3. **Imbalanced question categories**. The current version of the TCM-Ladder dataset exhibits noticeable imbalance across task types and knowledge subdomains. For instance, basic theoretical questions are overrepresented, while more complex or specialized areas such as clinical case reasoning, tongue image interpretation, and herbal property analysis are underrepresented. This imbalance may cause the models to overfit to high-frequency categories, leading to suboptimal generalization in low-resource tasks and ultimately affecting the fairness and representativeness of the overall evaluation. Future versions of the dataset should aim for more balanced coverage across task types and knowledge dimensions to ensure more comprehensive and challenging model assessment.

Despite the aforementioned limitations, TCM-Ladder, as the first multimodal dataset in the field of TCM, provides a foundational resource for the training and evaluation of multimodal TCM models. Evaluation results indicate that current TCM-specific large language models still exhibit substantial room for improvement and may lead to the following societal impacts.

1. **Facilitating the digitization and global dissemination of TCM**. By systematically organizing textual and visual modalities of TCM, the dataset contributes to the digitization of TCM knowledge.

2. **Enhancing education and clinical decision support**. This dataset provides a foundational resource for building intelligent TCM educational tools and diagnostic support systems. It may help address challenges such as the shortage of trained TCM professionals and the heavy reliance on experiential knowledge, ultimately supporting more standardized and scalable applications in clinical and educational settings.

3. **Privacy and fairness considerations**. Visual data such as tongue and herb images raise privacy concerns and require informed consent and anonymization. Additionally, the dataset may exhibit demographic imbalances (e.g., in gender, age, or region), which could introduce bias into trained

models. Future versions should incorporate fairness-aware data collection and annotation practices to mitigate these risks and support equitable downstream applications.

References

[1] Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, 2021.

[2] Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, KK Suhail, Sadhana Gonuguntla, et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific data*, 10(1):397, 2023.

[3] Harry Coppock, George Nicholson, Ivan Kiskin, Vasiliki Koutra, Kieran Baker, Jobie Budd, Richard Payne, Emma Karoune, David Hurley, Alexander Titcomb, Sabrina Egglestone, Ana Tendero Cañadas, Lorraine Butler, Radka Jersakova, Jonathon Mellor, Selina Patel, Tracey Thornley, Peter Diggle, Sylvia Richardson, Josef Packham, Björn W. Schuller, Davide Pigoli, Steven Gilmour, Stephen Roberts, and Chris Holmes. Audio-based ai classifiers show no evidence of improved covid-19 screening over simple symptoms checkers. *Nature Machine Intelligence*, 2024.

[4] BM Rocha, Dimitris Filos, Lea Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al. A respiratory sound database for the development of automated classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*, pages 33–37. Springer, 2018.

[5] Konstantia Zarkogianni, Edmund Dervakos, George Filandrianos, Theofanis Ganitidis, Vasiliki Gkatzou, Aikaterini Sakagianni, Raghu Raghavendra, CL Nikias, Giorgos Stamou, and Konstantina S Nikita. The smarty4covid dataset and knowledge base: a framework enabling interpretable analysis of audio signals. *arXiv preprint arXiv:2307.05096*, 2023.

[6] Elijah Moothedan, Micah Boyer, Stephanie Watts, Yassmeen Abdel-Aty, Satrajit Ghosh, Anaïs Rameau, Alexandros Sigaras, Olivier Elemento, Bridge2AI-Voice Consortium, and Yael Bensoussan. The bridge2ai-voice application: initial feasibility study of voice data acquisition through mobile health. *Frontiers in Digital Health*, 7:1514971, 2025.

[7] Ugo Cesari, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino, and Laura Verde. A new database of healthy and pathological voices. *Computers & Electrical Engineering*, 68:310–321, 2018.

[8] Patrick R Walden. Perceptual voice qualities database (pvqd): database characteristics. *Journal of Voice*, 36(6):875–e15, 2022.

[9] Jobie Budd, Kieran Baker, Emma Karoune, Harry Coppock, Selina Patel, Ana Tendero Cañadas, Alexander Titcomb, Richard Payne, David Hurley, Sabrina Egglestone, et al. A large-scale and pcr-referenced vocal audio dataset for covid-19. *arXiv preprint arXiv:2212.07738*, 2022.

[10] Raffaele Dubbioso, Myriam Spisto, Laura Verde, Valentina Virginia Iuzzolino, Gianmaria Senerchia, Elena Salvatore, Giuseppe De Pietro, Ivanoe De Falco, and Giovanna Sannino. Voice signals database of als patients with different dysarthria severity and healthy controls. *Scientific Data*, 11(1):800, 2024.

[11] Jobie Budd, Kieran Baker, Emma Karoune, Harry Coppock, Selina Patel, Richard Payne, Ana Tendero Cañadas, Alexander Titcomb, David Hurley, Sabrina Egglestone, et al. A large-scale and pcr-referenced vocal audio dataset for covid-19. *Scientific data*, 11(1):700, 2024.

[12] Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.

[13] Ivandro Sanches, Victor V Gomes, Carlos Caetano, Lizeth SB Cabrera, Vinicius H Cene, Thomas Beltrame, Wonkyu Lee, Sanghyun Baek, and Otávio AB Penatti. Mimic-bp: A curated dataset for blood pressure estimation. *Scientific Data*, 11(1):1233, 2024.

[14] Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal llms to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.

[15] Taha Samavati, Mahdi Farvardin, and Aboozar Ghaffari. Efficient deep learning-based estimation of the vital signs on smartphones. *arXiv preprint arXiv:2204.08989*, 2022.

[16] Ahsan Mehmood, Asma Sarouji, M Mahboob Ur Rahman, and Tareq Y Al-Naffouri. Your smartphone could act as a pulse-oximeter and as a single-lead ecg. *Scientific Reports*, 13(1):19277, 2023.

[17] Andrea Nemcova, Enikö Vargova, Radovan Smisek, Lucie Marsanova, Lukas Smital, and Martin Vitek. Brno university of technology smartphone ppg database (but ppg): Annotated dataset for ppg quality assessment and heart rate estimation. *BioMed Research International*, 2021(1):3453007, 2021.

[18] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023.

[19] Ping Yu, Kaitao Song, Fengchen He, Ming Chen, and Jianfeng Lu. Tcmd: A traditional chinese medicine qa dataset for evaluating large language models. *arXiv preprint arXiv:2406.04941*, 2024.

[20] Toyhom. Chinese-medical-dialogue-data. `https://github.com/Toyhom/Chinese-medical-dialogue-data`, 2025. Accessed: May 22, 2025.

[21] Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*, 2023.

[22] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[23] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[24] Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023.

[25] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376, 2024.

[26] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[27] IT Research Support Solutions Wiki. Getting started with hpc. `https://docs.itrss.umsystem.edu/pub/hpc/start`, 2025. Accessed: May 22, 2025.

[28] Shanghai kangqiao pharmaceutical co., ltd. `http://www.shkangqiao.com/`, 2025. Accessed: May 22, 2025.

[29] Jun Li, Pei Yuan, Xiaojuan Hu, Jingbin Huang, Longtao Cui, Ji Cui, Xuxiang Ma, Tao Jiang, Xinghua Yao, Jiacai Li, et al. A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. *Journal of biomedical informatics*, 115:103693, 2021.

[30] Jiacheng Xie, Congcong Jing, Ziyang Zhang, Jiatuo Xu, Ye Duan, and Dong Xu. Digital tongue image analyses for health assessment. *Medical Review*, 1(2):172–198, 2021.

[31] Qiujie Lv, Guanxing Chen, Haohuai He, Ziduo Yang, Lu Zhao, Kang Zhang, and Calvin Yu-Chian Chen. Tcmbank-the largest tcm database provides deep learning-based chinese-western medicine exclusion prediction. *Signal Transduction and Targeted Therapy*, 8(1):127, 2023.

[32] World Health Organization et al. *WHO international standard terminologies on traditional Chinese medicine*. World Health Organization, 2022.

[33] Xianwei Zhang, Peng Wu, Jiuming Cai, and Kun Wang. A contrastive study of chinese text segmentation tools in marketing notification texts. In *Journal of Physics: Conference Series*, volume 1302, page 022010. IOP Publishing, 2019.

[34] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[37] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[38] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[39] Hai-Yu Xu, Yan-Qiong Zhang, Zhen-Ming Liu, Tong Chen, Chuan-Yu Lv, Shi-Huan Tang, Xiao-Bo Zhang, Wei Zhang, Zhi-Yong Li, Rong-Rong Zhou, et al. Etcm: an encyclopaedia of traditional chinese medicine. *Nucleic acids research*, 47(D1):D976–D982, 2019.

[40] Kai Gao, Liu Liu, Shuangshuang Lei, Zhinong Li, Peipei Huo, Zhihao Wang, Lei Dong, Wenxin Deng, Dechao Bu, Xiaoxi Zeng, et al. Herb 2.0: an updated database integrating clinical and experimental evidence for traditional chinese medicine. *Nucleic Acids Research*, 53(D1):D1404–D1414, 2025.

[41] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[42] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[43] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.