

ShizhenGPT: Towards Multimodal LLMs for Traditional Chinese Medicine

Junying Chen^{1†}, Zhenyang Cai^{1†}, Zhiheng Liu¹, Yunjin Yang¹, Rongsheng Wang¹
Qingying Xiao³, Xiangyi Feng², Zhan Su⁶, Jing Guo⁵

Xiang Wan⁴, Guangjun Yu^{1,3}, Haizhou Li¹, Benyou Wang^{1,4*}

¹The Chinese University of Hong Kong, Shenzhen ²Freedom AI

³National Health Data Institute, Shenzhen ⁴Shenzhen Research Institute of Big Data

⁵Guangdong University of Technology ⁶Halmstad University, Sweden

wangbenyou@cuhk.edu.cn

Abstract

Despite the success of large language models (LLMs), their potential in **Traditional Chinese Medicine (TCM)** remains underexplored due to scarce high-quality data and the multi-modal nature of TCM diagnostics, which involve looking, listening, smelling, and pulse-taking. These modalities challenge conventional LLMs. To address these challenges, we present **ShizhenGPT**, the first multimodal LLM tailored for TCM. To overcome data scarcity, we curate the largest TCM dataset to date, comprising 100GB+ of text and 200GB+ of multimodal data, including 1.2M images, 200 hours of audio, and physiological signals. ShizhenGPT is pretrained and instruction-tuned to achieve deep TCM knowledge and multimodal reasoning. For evaluation, we collect recent national TCM qualification exams and build a visual benchmark for Medicinal Recognition and Visual Diagnosis. Experiments demonstrate that ShizhenGPT outperforms comparable-scale LLMs and competes with larger proprietary models. Moreover, it leads in TCM visual understanding among existing multimodal LLMs and shows promising results across modalities such as sound, pulse, electronic-nose signals, and vision, paving the way toward more holistic multimodal perception in TCM. We hope that the datasets, models, and methods of this work will inspire further exploration in this field. Code, models, and released data are available at <https://github.com/FreedomIntelligence/ShizhenGPT>.

1 Introduction

Large language models (LLMs) like GPT-4 have demonstrated impressive capabilities across general and specialized domains (Achiam et al., 2023; Dubey et al., 2024; Guo et al., 2025). Meanwhile, Traditional Chinese Medicine (TCM), a medical system with thousands of years of history and on-going relevance for hundreds of millions (Shi and

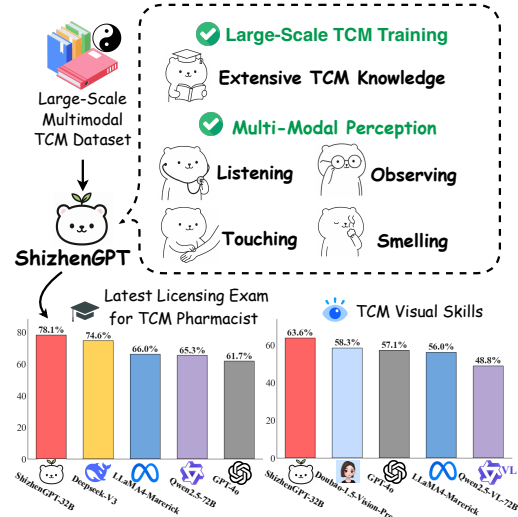


Figure 1: Key Capabilities of ShizhenGPT, a multimodal LLM for Traditional Chinese Medicine (TCM).

Zhou, 2007; Zhao et al., 2015), remains largely absent from recent AI developments. TCM, with its intricate theoretical framework, nuanced diagnostic reasoning, and enduring clinical relevance (Zhu and Wang, 2010), presents both a challenge and an opportunity for large language models (Hua et al., 2024). Empowering these models to understand and reason about TCM could greatly advance clinical decision-making, medical education, and preservation of traditional medical knowledge (Yao et al., 2019; Wang et al., 2020).

However, applying LLMs to TCM presents two key challenges. First, there is a significant scarcity of high-quality TCM data. For example, most existing TCM-specific LLMs are trained on less than 1GB of text (Dai et al., 2024; Yang et al., 2024c; Wei et al., 2025; Hua et al., 2024). This is substantially smaller than the datasets used in domains like modern medicine (Zhang et al., 2024b; Chen et al., 2024c), and particularly insufficient given the complexity and depth of TCM theory. Moreover, the multimodal datasets necessary for

[†] Equal contribution. * Corresponding author.

Model	Domain Expertise	Multimodal Capability					Clinical Application
		Text 	Vision 	Sound 	Smell 	Pulse 	
TCM-specific LLMs	Limited by small-scale training data.	✓	✗	✗	✗	✗	Limited to text-only scenarios.
General-purpose LLMs	General-domain training, limited TCM specialization.	✓	○ Partial	○ Partial	✗	✗	Only partial support for vision/sound.
ShizhenGPT 	Trained on 15B+ tokens and 200GB+ TCM data; high TCM proficiency.	✓	✓	✓	✓	✓	Toward Four Diagnostic Methods.

Table 1: Comparison of ShizhenGPT and existing LLMs across key dimensions.

training TCM-oriented multimodal LLMs are currently non-existent. Second, TCM fundamentally depends on multimodal diagnostics. Rooted in the "Four Diagnostic Methods", it depends on observing the tongue and other visual cues, listening to the voice and breath, detecting odors, and feeling the pulse (Dong et al., 2022; Tian et al., 2023). These rich sensory modalities are central to TCM and lie well beyond the capabilities of text-only models.

To address this, we propose **ShizhenGPT**, the first multimodal LLM tailored for Traditional Chinese Medicine. To mitigate data scarcity, we built the largest TCM dataset to date, comprising over 100GB of text from 3,256 TCM-specific books and online sources. In addition, we collected over 200GB of multimodal data, including 1.2 million annotated images, 200+ hours of audio, and diverse physiological signals such as pulse, smell, and electrocardiograms (ECG). Leveraging this extensive dataset, ShizhenGPT is trained through domain-specific pretraining and instruction tuning, enabling it to acquire deep TCM knowledge and multimodal reasoning. The resulting model ShizhenGPT integrates deep domain expertise with perceptual capabilities aligned with real-world TCM diagnostic practices, including observation, listening, smelling, and pulse-taking.

To evaluate TCM-specific capabilities, we collected recent qualification exams, including three national licensing exams and two postgraduate entrance exams in TCM. All questions were released within the past year to ensure freshness. For visual evaluation, we built a benchmark using TCM atlases, containing 7k image-related questions focused on medicinal recognition and visual diagnosis. Results show that ShizhenGPT outperforms existing LLMs of similar scale and rivals much larger proprietary models, despite having only 32B parameters. In visual TCM tasks, it leads among current multimodal LLMs. Expert evaluations also indicate a higher preference for ShizhenGPT’s re-

sponses. Additionally, across seven signal datasets, ShizhenGPT shows promising preliminary results on sound, pulse, smell, and ECG tasks using a unified input interface.

We present a TCM-specific multimodal LLM capable of perceiving smell, sound, vision, and pulse. This expands the capabilities of the model beyond text-based interaction, allowing it to process tongue images, pulse signals, and breathing sounds within a unified interface. This moves toward more realistic, holistic medical AI systems. We hope our data, model, and approach will inspire further research in TCM.

Our contributions are as follows:

- We release the largest TCM dataset to date, covering extensive text and multimodal data.
- We introduce ShizhenGPT, the first multimodal LLM tailored for TCM, capable of understanding images, sounds, smells, pulse, and more. It achieves strong performance in TCM expertise and leads in visual diagnostic tasks.
- We release a comprehensive benchmark for TCM across textual and multimodal tasks, enabling systematic evaluation of LLMs in TCM.

2 Background and Motivation

Traditional Chinese Medicine TCM is one of the world’s oldest and most influential medical systems, serving over 500 million patients annually in China and gaining global recognition (Cheung, 2011; Liu et al., 2015; Zhao et al., 2023). Unlike modern medicine, TCM relies on the "Four Diagnostic Methods": **looking, listening/smelling, questioning, and pulse-taking** (Dong et al., 2022; Tian et al., 2023). These methods collect visual, auditory, verbal, and tactile information to form a holistic understanding of the patient. Each captures distinct biomedical signals, making TCM inherently multimodal and posing unique challenges for AI systems seeking to replicate its diagnostic rea-

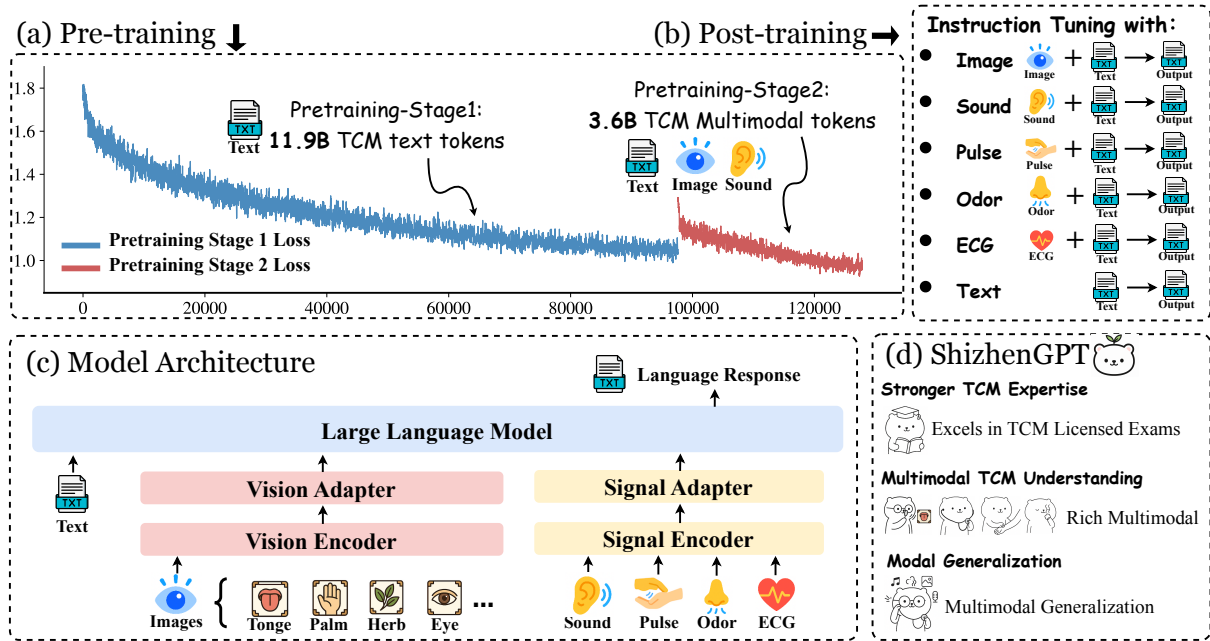


Figure 2: Overview of ShizhenGPT. (a) Pre-training process, with the loss curve of *ShizhenGPT-7B*. (b) Post-training process with multimodal instruction tuning. (c) Model architecture. (d) Demonstration of ShizhenGPT’s capabilities.

soning.

Motivation for a Multimodal TCM LLM TCM’s complexity and its reliance on diverse sensory modalities pose significant challenges for LLMs. As shown in Table 1, current TCM-specific models suffer from limited training data and lack multimodal perception. General-purpose models, while powerful, are not tailored to TCM and offer only partial support for modalities like vision or audio. To address these gaps, we propose **ShizhenGPT**, a multimodal LLM trained on extensive TCM data in diverse modalities. By integrating perception and reasoning across all four diagnostic methods, ShizhenGPT might enable more faithful, practical applications in real-world clinical settings.

3 ShizhenGPT

This section describes **ShizhenGPT**, including its architecture and two-phase training: (1) *pre-training* to inject TCM-specific knowledge, and (2) *post-training* for real-world tasks.

3.1 Model Architecture

As shown in Figure 2(c), ShizhenGPT consists of three main components: an LLM backbone, a vision encoder, and a signal encoder.

LLM Backbone The LLM serves as the core reasoning engine. It processes inputs from mul-

tiple modalities and generates responses. We use Qwen2.5-7B and Qwen2.5-32B (Yang et al., 2024a) as the base LLMs.

Vision Encoder For visual inputs, we initialize the *Vision Encoder* from Qwen2.5-VL (Bai et al., 2025), which supports high-resolution images using 2D-RoPE and window attention. Visual patches extracted by the encoder are grouped (four adjacent patches), concatenated, and mapped to the LLM’s embedding space via a two-layer MLP (*Vision Adapter*).

Signal Encoder For continuous signals (e.g., voice, pulse, smell), we use Whisper-large-v3 (Radford et al., 2023) as the initial *Signal Encoder*. Non-audio signals are first converted into waveform representations via linear interpolation. All waveforms are resampled to 16kHz and transformed into 128-channel mel-spectrograms using a 25ms window and 10ms hop size. A stride-2 pooling layer reduces temporal resolution, and a one-layer MLP (*Signal Adapter*) projects the features into the embedding space of LLM. Although pulse signals are low-frequency, the resulting 40ms-per-token resolution retains sufficient temporal granularity to capture the slow-varying patterns essential for TCM. Appendix C.3 compares alternative signal representations.

Pre-training Dataset	Description	# Tokens	Storage
 Text Datasets (Pretraining Stage 1 & Stage 2)			
 TCM Web Corpus	A high-quality TCM corpus collected from Common Crawl (2017–2023) and WeChat public articles (a major content channel in China).	5.33B	21.2 GB
 TCM Book Corpus	A cleaned corpus of 3,256 TCM textbooks. (2 epochs)	0.96B	3.8 GB
 General Text Corpus	General-domain English corpus from FineWeb-edu. See Appendix D for details.	5.63B	29.2 GB
 Image-Text Dataset (Pretraining Stage 2)			
 TCM Web Interleaved Data	Interleaved text-image data from the TCM web corpus. (# Images: 1,143,954)	0.87B	140.7 GB
 TCM Book Interleaved Data	Richly illustrated interleaved text-image data from 306 TCM books. (# Images: 51,619 ; 2 epochs)	0.14B	17.6 GB
 TCM Synthesized Image-Text Data	TCM image-text pairs generated from images and their context using multi-modal LLMs. (# Images: 146,635 ; 2 epochs)	0.16B	40.6 GB
 General Image-Text Data	General-domain image-text datasets, such as ShareGPT-4V. See Appendix D for details. (# Images: 1,851,321)	0.69B	93.29 GB
 Audio-Text Dataset (Pretraining Stage 2)			
 TCM-related Audio-Text Data	Chinese audio-text pairs in the domain of traditional Chinese medicine. (# Audios: 58,462 ; Duration: 238.67 hours)	0.03B	75.72 GB
 General Audio-Text Data	English audio-text pairs from general domains. See Appendix D. (# Audios: 282,884 ; Duration: 353.04 hours)	0.04B	111.6 GB

Table 2: Overview of **Pre-training Datasets**. These data are utilized for a single training epoch unless otherwise specified. # indicates quantity. **B** stands for billion.

3.2 Pre-training

We begin with pre-training on large-scale TCM data to acquire domain expertise. We design a two-stage pipeline: the first stage infuses knowledge from extensive TCM text, while the second introduces multimodal alignment through image-text and audio-text data.

Pretraining Datasets As summarized in Table 2, the pretraining corpus includes three components:

- **TCM Text Corpus** We collect TCM text from 3,256 TCM books and online sources. The books are converted into interleaved text-image corpora and cleaned to yield 3.8GB of *book corpus*. For online sources, we use a 30K-term TCM lexicon to extract TCM documents from Common Crawl (2017–2023) and WeChat public articles, totaling 96.4GB. A two-step filtering process, involving quality scoring using a classifier and semantic deduplication, produces a final 21.2GB high-quality *web corpus*.
- **TCM Image-Text Data** We extract 17.6GB of interleaved image-text data from TCM books (51K images). From 5 million images sourced from WeChat articles, we filtered high-quality TCM images using an image classifier. This process yielded 140.7GB of web-based image-text data (1M images). To improve alignment, we follow HuatuoGPT-Vision and use a multimodal LLM to generate descriptive captions

based on image and context, producing 40.6GB of synthetic pairs (146K images).

- **TCM Audio-Text Data** We extract 60K doctor-patient dialogues from the Huatuo-26M dataset (Wang et al., 2025) and synthesize audio via a high-fidelity TTS system (Du et al., 2024), yielding 58K utterance-aligned audio-text pairs.

Further details are provided in Appendix E.

Pre-training Strategy We adopt a two-stage pre-training strategy. In **Stage 1**, it focuses on text-only learning with 11.9B tokens, including 6.3B from TCM corpora and 5.6B from general corpora (see Appendix D for details) to maintain foundational capabilities. In **Stage 2**, it uses 3.8B tokens of multimodal data, including TCM and general image-text and audio-text data, along with 1.8B resampled textual tokens from Stage 1 to preserve grounding. In both stages, inputs are packed into 4096-token sequences, and loss is computed only over textual tokens. Pre-training runs for 1 epoch with a 5e-5 learning rate using full-parameter tuning. Details are provided in Appendix F.

3.3 Post-training

After pre-training, the model acquires foundational TCM knowledge and multimodal abilities. Post-training aligns the model for instruction-following, extends its capabilities to downstream tasks, and

Instruction-Tuning Dataset	Description	# Item
 Text Instruction Data		
 TCM Instruction	A dialogue instruction dataset centered on real-world TCM problems.	83,629
 Vision Instruction Data (Requires Vision Encoder)		
 TCM Vision Instruction	A multimodal instruction dataset composed of TCM-related image-text pairs. (# Images: 70,638)	65,033
   Signal Instruction Data (Requires Signal Encoder)		
 TCM Speech Instruction	Speech-to-text instruction data in the TCM domain.	57,957
 Cough Sound Dataset	An instruction dataset focused on cough sounds for medical diagnosis.	456
 Heartbeat Sound Dataset	An instruction dataset focused on heartbeat sounds for medical diagnosis.	189
 Pulse Dataset	An instruction dataset focused on pulse diagnosis in TCM applications.	4,101
 Smell Dataset	Dataset related to olfactory input for medical diagnosis.	672
 ECG Dataset	Electrocardiogram instruction dataset for medical diagnosis.	99,195

Table 3: Overview of **Instruction-Tuning Datasets**. Signal instructions are detailed in Appendix J.

enables adaptation to additional modalities such as sound and smell, which require low learning complexity but suffer from limited data.

Instruction Dataset As summarized in Table 3, instruction tuning data spans four categories:

- **TCM Text Instructions** We curate 92K TCM questions from doctor-patient dialogues (Wang et al., 2025) and medical verifiable questions (Chen et al., 2025a). Responses were distilled using DeepSeek-R1 and validated with DeepSeek-V3 based on reference answers or doctor responses, resulting in 83K reliable Chinese instruction pairs.
- **Vision Instructions** We select TCM-related images with detailed captions from books and web sources. GPT-4o is used to generate paired visual questions and answers based on image content and context, producing 65K high-quality vision instruction pairs. Appendix H.1 reports a spot check of this synthetic supervision.
- **Audio Instructions** For speech audio, we convert 20K text questions into TTS speech and pair them with original answers. An additional 20K transcription-style instructions are created. For TCM-specific audio (e.g., coughs, heart sounds), classification tasks are reformulated into instruction format (e.g., “Listen and determine if the patient has COVID-19.”).
- **Physiological Signal Instructions** Non-audio signals (e.g., pulse, smell, ECG) are converted to waveform format and tagged with modality-specific tokens (e.g., <Smell>, <Pulse>). The pregnancy-related pulse data is constructed in-

house, while the other data are obtained from public or authorized datasets. All tasks are transformed into instruction format using a unified signal-processing pipeline to ensure extensibility.

The instruction and signal datasets are detailed in Appendices H and J, respectively.

Instruction Fine-tuning We fine-tune the TCM-pretrained model on all modality-specific instruction datasets in Table 3 using full-parameter tuning. Instructions follow the *Qwen2.5* chat template, and loss is computed only on response tokens. The model is trained for 3 epochs with a learning rate of $5e-6$. See Appendix F for details. This stage produces the final model, **ShizhenGPT**.

4 Experimental Setup

Model Training We train two versions: **ShizhenGPT-7B** and **ShizhenGPT-32B**, based on *Qwen2.5-7B* and *Qwen2.5-32B* backbones, respectively. Both use the same vision and signal encoders, and are trained with identical settings across all stages. Experiments are conducted on four DGX nodes, each with 8 A100 GPUs.

Baselines We compare ShizhenGPT with following baselines: (1) **Vanilla LLMs**: We evaluate mainstream general-purpose models, including LLaMA (Dubey et al., 2024), Qwen, GLM-4 (GLM et al., 2024), DeepSeek (Liu et al., 2024a), GPT-4o (2024-11-20) (Achiam et al., 2023), GPT-5 (gpt-5.1-2025-11-13), Doubao (Doubao Team, 2025), etc. TCM-specific LLMs include TCMChat-600K-7B (Dai et al., 2024), Zhongjing-7B (Yang et al., 2024c), Biancang-7B (Wei et al., 2025), and

Model	2024 TCM Pharmacist	2024 TCM Physician	2024 TCM Assistant Physician	2024 TCM Graduate Entrance	2025 TCM Graduate Entrance	Avg.
🔴 Proprietary or 100B+ LLMs						
🔴 GPT-4o	61.7	56.2	66.7	58.7	55.0	59.6
🔴 GPT-5	63.2	57.2	69.2	59.5	52.5	60.3
🔴 LLaMA-4-Maverick-400B	66.0	67.2	76.6	63.3	59.5	66.5
🔴 DeepSeek-V3	74.6	74.8	80.8	72.4	66.7	73.9
🔴 Gemini-3-Pro	86.0	77.0	84.8	79.8	74.8	80.5
🔴 Doubao-1.5-Pro-32k	81.9	79.7	86.2	81.5	73.7	80.6
🔴 DeepSeek-R1-671B	85.0	82.4	86.2	83.2	83.3	84.0
<15B LLMs						
Llama-3.1-8B-Instruct	36.7	36.9	40.4	33.4	26.1	34.7
🌀 TCMChat-600k-7B	42.0	49.5	46.0	41.6	38.5	43.5
🌀 Zhongjing-7B	45.5	47.8	46.4	38.4	45.6	44.7
🌀 Biancang-7B	57.3	57.9	66.2	43.5	37.5	52.5
GLM4-9B-Chat	54.6	56.2	61.6	54.0	51.9	55.7
🌀 Lingdan-13B-PR	64.4	65.9	78.4	45.6	41.1	59.1
DeepSeek-R1-Distill-14B	62.2	60.8	70.5	58.4	55.7	61.5
Qwen2.5-7B-Instruct	65.3	63.2	76.0	59.0	54.0	63.5
Qwen3-8B	66.3	64.9	78.3	57.7	52.3	63.9
InternLM2.5-7B-chat	69.8	68.5	79.7	58.5	48.9	65.1
🌀 ShizhenGPT-7B (Ours)	70.1	76.7	79.4	68.8	69.3	72.9
15B to 100B LLMs						
Gemma-3-27B-it	42.1	40.9	44.9	35.1	28.1	38.2
Llama-3.1-70B-Instruct	50.6	52.8	65.0	42.5	39.9	50.2
InternLM2.5-20B-Chat	66.5	66.1	73.8	57.9	53.8	63.6
GLM-4-32B	67.9	65.6	80.9	64.9	62.6	68.4
DeepSeek-R1-Distill-32B	70.6	67.7	78.4	65.3	61.2	68.6
Qwen3-32B	73.3	68.5	80.0	65.6	63.8	70.2
Qwen2.5-32B-Instruct	76.1	69.2	78.3	65.5	61.8	70.2
Qwen2.5-72B-Instruct	77.2	70.6	81.0	63.3	59.9	70.4
🌀 ShizhenGPT-32B (Ours)	79.2	78.4	84.9	77.5	70.7	78.1

Table 4: Results of LLMs on recent TCM exams, including the 2024 Licensed TCM Pharmacist Exam, 2024 Licensed TCM Physician Exam, 2024 Licensed TCM Assistant Physician Exam, and two TCM Postgraduate Entrance Exams (2024-2025). 🔴 denotes proprietary or 100B+ LLMs; 🌀 denotes TCM-specific LLMs.

Lingdan-13B-PR (Hua et al., 2024). (2) **Multi-modal LLMs (MLLMs)**: For vision understanding, we include Gemini (Team et al., 2023), LLaVA (Liu et al., 2023), Qwen2.5-VL (Bai et al., 2025), InternVL (Chen et al., 2024e), Janus-Pro (Chen et al., 2025b), Kimi-VL (Team et al., 2025) and HuatuoGPT-Vision (Chen et al., 2024b); for audio understanding, we consider Qwen2-Audio (Chu et al., 2024), Llama-Omni (Fang et al., 2025), and Moshi (Défossez et al., 2024).

Benchmarks To evaluate the TCM capabilities, we construct a multimodal benchmark suite covering text, vision, signals, and human evaluation. (1) **Text Benchmarks**: We collect recent national TCM exams in China, including three licensing exams (for pharmacists, physicians, and assistant physicians) and two postgraduate entrance exams (2024–2025). (2) **Vision Benchmark**: We compile 7,204 multiple-choice questions from eight TCM atlases, explicitly excluded from training, covering

eight subfields in medicinal recognition and visual diagnosis. See Appendix K for details. (3) **Human Benchmark**: We collect 90 real-world TCM questions and evaluate model responses, with licensed TCM doctors selecting preferred answers (see Appendix L). (4) **Signal Benchmarks**: We use the test set of the dataset for smell, auscultation, pulse, and ECG signals, along with a custom pregnancy detection dataset using pulse data. Additional split details and stronger signal baselines are reported in Appendix C.3.

5 Results

Evaluation of TCM Expertise Table 4 reports the performance of various models on recent National TCM Licensing exams and TCM postgraduate entrance exams. For fair comparison, models are grouped by parameter size. In the small-scale category (<15B), **ShizhenGPT-7B** achieves the highest average score, outperform-

Model	TCM Medicinal Recognition			TCM Visual Diagnosis					Avg
	TCM Patent	TCM Material	TCM Herb	Holism	Tongue	Palm	Tuina	Eye	
Multimodal LLMs									
LLaVA-1.5-7B	25.6	26.9	27.7	25.8	27.0	19.2	16.9	23.3	24.1
LLaVA-Med	26.2	35.4	23.1	24.8	28.7	26.5	43.5	38.4	30.8
DeepSeek-VL2	30.0	39.1	26.6	28.2	32.4	30.8	47.2	42.0	34.5
MedGemma-4B-it	30.2	40.6	28.0	46.7	49.9	41.5	38.3	38.1	39.2
Gemma-3-27B-it	32.6	40.1	31.8	45.8	48.8	46.2	38.1	45.3	41.1
Llama-3.2-11B-Vision	33.1	43.5	31.1	50.0	53.0	44.5	41.8	41.3	42.3
Janus-Pro-7B	32.5	45.1	33.0	45.7	48.6	45.3	45.8	50.8	43.4
HuatuoGPT-Vision-7B	38.4	53.0	36.2	48.9	52.6	44.9	50.4	49.0	46.7
InternVL3-8B	38.8	58.6	39.0	58.5	54.8	45.2	53.3	47.1	49.4
Qwen2.5-VL-32B-Instruct	39.6	60.3	46.0	61.2	56.4	54.8	51.4	49.2	52.4
Kimi-VL-Instruct-16B	47.1	64.6	46.4	60.5	53.0	52.3	51.9	51.9	53.5
InternVL3-38B	46.1	65.8	48.8	58.2	57.3	49.4	54.8	49.7	53.7
Qwen2.5-Omni-7B	43.8	61.8	51.2	60.8	54.5	53.5	59.5	52.0	54.6
🔴 LLaMA-4-Maverick-400B	46.0	67.8	52.0	61.0	62.5	55.0	54.8	49.2	56.0
🔴 GPT-4o	48.0	70.6	54.5	63.2	57.9	57.4	57.5	47.8	57.1
🔴 GPT-5	52.3	77.3	53.3	61.7	56.3	59.0	56.7	49.7	58.2
🔴 Doubao-1.5-Vision-Pro	52.0	79.3	64.0	59.0	62.8	45.7	58.2	45.2	58.3
🔴 Gemini-3-Pro	52.3	74.7	62.3	63.7	60.2	61.7	62.0	54.5	61.4
ShizhenGPT (Ours)									
ShizhenGPT-7B	47.4	61.7	49.2	70.6	65.5	63.7	64.1	58.3	60.1
ShizhenGPT-32B	49.5	74.2	54.7	71.5	66.5	60.9	65.9	65.3	63.6

Table 5: Results on the **TCM-Vision Benchmark**, evaluating visual understanding in TCM.  denotes proprietary or 100B+ multimodal LLMs. **bold** indicates the best score; underlines marks the second-best.

ing Qwen2.5-7B-Instruct by +9.4 points despite sharing the same backbone. It also surpasses larger models such as DeepSeek-R1-Distill-14B, and leads all compared TCM-specific models in this range. In the mid-scale range (15-100B), **ShizhenGPT-32B** achieves the strongest results among open-source LLMs, with an average score of 78.1. This exceeds multiple 70B+ models. Although it slightly lags behind ultra-large proprietary models like DeepSeek-R1 (671B) and Doubao-1.5-Pro-32K, it demonstrates highly competitive performance with significantly fewer parameters (78.1 vs. 84.0).

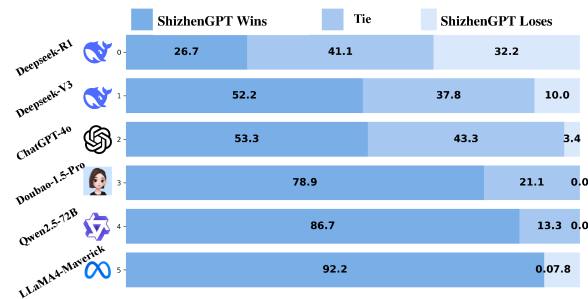


Figure 3: Results of the human evaluation. ShizhenGPT refers to *ShizhenGPT-32B*. "Win/Tie/Loss" indicates the proportion of expert preferences for the model responses (Details are in Appendix L).

Human Evaluation To further assess TCM capabilities, we conduct an expert evaluation with licensed TCM practitioners. Following (Zhang et al., 2023), we sample 90 real-world TCM queries. For each, our model and a baseline generate responses, which are then compared by experts. ShizhenGPT-32B outperforms all baseline models except DeepSeek-R1, including strong contenders like Doubao-1.5-Pro and DeepSeek-V3. The performance gap with DeepSeek-R1 is minimal. Given the substantial inference cost of DeepSeek-R1 with 671B parameters, ShizhenGPT-32B offers a compelling trade-off between model size and clinical performance.

Evaluation of Visual Capabilities Table 5 presents results on the TCM visual benchmark, which evaluates the ability to recognize medicinal herbs and interpret diagnostic visuals (e.g., tongue, palm). Proprietary models like GPT-4o demonstrate stronger visual reasoning than open-source models such as Qwen2.5-VL. Among baselines, Gemini-3-Pro, a proprietary MLLM, achieves the strongest baseline score of 61.4. **ShizhenGPT** surpasses all baselines, achieving a new SOTA with 63.6 using only 32B parameters. Notably, even the 7B version performs competitively, highlighting the effectiveness of domain-specific multimodal

training for specialized visual understanding in TCM. Table 5 reports seed-0 results. Across five inference seeds, ShizhenGPT-7B and ShizhenGPT-32B average 60.2 ± 0.74 and 63.3 ± 0.91 , respectively. Additional benchmark checks are reported in Appendix K.

Model	General Audio LLMs Benchmarks		
	LlamaQuestion	Speech TriviaQA	Speech Web Questions
Llama-Omni	45.3	22.9	10.7
Moshi	43.7	23.8	16.7
Qwen2-Audio	60.0	30.4	24.0
ShizhenGPT-7B	64.0	27.3	25.4

Table 6: Results on general speech tasks.

Evaluation of General Audio Capabilities

Since ShizhenGPT is equipped with audio perception, we evaluate its general speech understanding ability on standard audio benchmarks. As shown in Table 6, ShizhenGPT demonstrates strong performance, comparable to *Qwen2-Audio-Instruction*. These results indicate that ShizhenGPT is not only compatible with various signal modalities but also maintains strong audio comprehension.

Modality	Task	Random Baseline	ShizhenGPT (7B)
Pulse	Pregnancy Detection from Pulse Signal	50.0	80.5
	Disease Classification from Pulse Signal	11.1	20.6
Smell	Disease Classification from E-nose Signal	14.3	48.8
Sound	COVID-19 Detection from Cough Sound	50.0	58.7
	Cardiac Abnormality Detection from Heart Sound	50.0	62.9
ECG	ECG Abnormality Detection	50.0	71.5
	ECG Heart Beat Classification	20.0	83.1

Table 7: Performance on various signal modalities. All tasks are single-modality classification. Details and additional baselines are in Appendix C.3.

Evaluation of Signal Modalities Table 7 shows the model’s performance across signal modalities, such as smell, ECG, and pulse. The model consistently outperforms random baselines, demonstrating its ability to process diverse sensory inputs through a unified interface. Five-fold cross-validation and stronger encoder baselines are reported in Appendix C.3. The pregnancy-pulse experiment evaluates signal representation and should not be interpreted as a clinical diagnostic test.

Ablation Study on Pre-training We conduct an ablation study to evaluate the role of TCM-specific pre-training. As shown in Table 8, extensive TCM pre-training substantially enhances TCM expertise

Model	TCM Expertise		TCM Visual Capability	
	2024 TCM Pharmacist	2024 TCM Physician	Medicinal Recognition	Visual Diagnosis
ShizhenGPT-7B	70.1	76.7	52.7	64.5
w/o Pretraining	66.9 (↓ 3.2)	68.8 (↓ 7.9)	43.7 (↓ 9.0)	59.3 (↓ 5.2)
ShizhenGPT-32B	79.2	78.4	59.5	66.0
w/o Pretraining	76.9 (↓ 2.3)	72.5 (↓ 5.9)	51.3 (↓ 8.2)	62.7 (↓ 3.3)

Table 8: Ablation study results on ShizhenGPT pre-training. Performance is evaluated across two TCM qualification exams and two TCM-specific visual benchmarks.

and visual understanding capabilities. This confirms that large-scale TCM training is effective for TCM-specific models.

Qualitative Examples Appendix N provides separate audio and image examples. They illustrate model behavior but are not evidence of clinical safety or patient-level multimodal fusion.

6 Conclusion

In this work, we introduce **ShizhenGPT**, the first multimodal LLM tailored for TCM, along with an extensive TCM dataset. Experiments show that ShizhenGPT surpasses existing LLMs of a similar scale. It leads in visual TCM tasks and shows promising results on sound, pulse, and smell signals. This extends its capabilities beyond text and enables tongue images, pulse signals, and breath sounds to be processed within a unified interface. This work moves toward more holistic medical AI and aims to inspire further integration of TCM and AI.

Limitations

- **Insufficient TCM Signal Data** Despite the large-scale collection of text and image data, high-quality signal data (e.g., smell, pulse) remains scarce. Most are sourced from limited public datasets, constraining the model’s ability to fully develop and generalize in these modalities.
- **Lack of Real-World Clinical Testing** Although expert evaluations were conducted, ShizhenGPT has not been tested in real clinical scenarios. The absence of patient-level feedback limits our understanding of its practical effectiveness and safety.
- **Incomplete Modal Coverage** Current modalities include vision, sound, and selected physiological signals, but other modals like tactile

sensation are still missing.

- **Synthetic Supervision Quality** Text instructions are filtered against reference answers, and an LLM-based spot check finds high grounding consistency for sampled visual QA and captions (Appendix H.1). This is not a substitute for verification by TCM practitioners, and factual noise may remain.
- **No Demonstrated Cross-modal Synergy** The signal ablations test whether one model retains separate capabilities under joint tuning. They do not establish synchronized patient-level fusion. Tactile examination is also not represented.

Despite these gaps, we hope this work will inspire deeper collaboration between computer scientists and medical practitioners, especially in building richer multimodal datasets to advance AI in Traditional Chinese Medicine.

Potential Risks

As a multimodal LLM for Traditional Chinese Medicine, the current ShizhenGPT model may still generate content that contains hallucinations or factual inaccuracies. As such, it is intended solely for scientific research and is not suitable for deployment in real-world applications. To mitigate potential risks, we impose strict limitations on its use: the model must not be applied in clinical practice or any industry-related scenarios where inaccuracies could lead to adverse outcomes. We strongly emphasize the ethical responsibility of users to adhere to these restrictions, in order to safeguard both the safety and integrity of downstream applications.

Acknowledgments

This work was supported by the Shenzhen Loop Area Institute (2026SFP004), the National Natural Science Foundation of China (72495131), its Tianyuan Fund for Mathematics (12326608), and the International Science and Technology Cooperation Center, Ministry of Science and Technology of China (2024YFE0203000). Support from the Shenzhen Medical Academy of Research and Translation (SMART) included the Shenzhen Medical Research Fund (C2406002 and B2503005) and the Major Frontier Exploration Program (C10120250085). Additional support came from the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence

(2023B1212010001), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2023ZT10X044), the Shenzhen Science and Technology Program (JCYJ20220818103001002 and ZDSYS20230626091302006), the Shenzhen Doctoral Startup Funding (RCBS20221008093330065), the Shenzhen Stability Science Program 2023, and the Shenzhen Key Laboratory of Multi-Modal Cognitive Computing.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.
- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [Disc-medllm: Bridging general large language models and real-world medical consultation](#). *Preprint*, arXiv:2308.14346.
- Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. 1995. [Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet](#). *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025a. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573. Association for Computational Linguistics.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024b. [Towards injecting medical visual knowledge into multimodal LLMs at scale](#).

- In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370. Association for Computational Linguistics.
- Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024c. [HuatuoGPT-II, one-stage training for medical adaptation of LLMs](#). In *First Conference on Language Modeling*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024d. [Sharegpt4v: Improving large multi-modal models with better captions](#). In *European Conference on Computer Vision*, pages 370–387. Springer.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024e. [Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.
- Felix Cheung. 2011. [Tcm: made in china](#). *Nature*, 480(7378):S82–S83.
- Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42—evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Yizheng Dai, Xin Shao, Jinlu Zhang, Yulong Chen, Qian Chen, Jie Liao, Fei Chi, Junhua Zhang, and Xiaohui Fan. 2024. [Tcmchat: A generative large language model for traditional chinese medicine](#). *Pharmacological Research*, 210:107530.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.
- Sui Dong, Lei Zhang, and Fei Yang. 2022. [Data-driven based four examinations in tcm: a survey](#). *Digital Chinese Medicine*, 5(4):377–385.
- Xin Dong, Chenxi Zhao, Xinpeng Song, Lei Zhang, Yu Liu, Jun Wu, Yiran Xu, Ning Xu, Jialing Liu, Haibin Yu, et al. 2024. [Presrecst: a novel herbal prescription recommendation algorithm for real-world patients with integration of syndrome differentiation and treatment planning](#). *Journal of the American Medical Informatics Association*, 31(6):1268–1279.
- Doubao Team. 2025. [Doubao-1.5-Pro](#). ByteDance Seed model release. Released: 2025-01-22; accessed: 2026-08-29.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. [LLaMA-Omni: Seamless speech interaction with large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Chaoxun Guo, Zhixing Jiang, Haoze He, Yining Liao, and David Zhang. 2022. [Wrist pulse signal acquisition and analysis for disease diagnosis: A review](#). *Computers in Biology and Medicine*, 143:105312.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.
- Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Linshan Wu, Ngai Shing Cheng, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, and 8 others. 2026. [Towards generalizable AI in medicine via generalist–specialist collaboration](#). *Nature Biomedical Engineering*.
- Rui Hua, Xin Dong, Yu Wei, Zixin Shu, Pengcheng Yang, Yunhui Hu, Shuiping Zhou, He Sun, Kaijing Yan, Xijun Yan, et al. 2024. [Lingdan: enhancing](#)

- encoding of traditional chinese medicine knowledge for clinical reasoning tasks with large language models. *Journal of the American Medical Informatics Association*, 31(9):2019–2029.
- Yongzhe Jia, Xiangyu Ji, Xin Wang, Heyi Zhang, Zhaopeng Meng, Junhua Zhang, Zhe Chen, Pengwei Zhuang, Dawei Xu, Wenbin Guo, Yuting Yan, and Jianguo Wei. 2025. [Qibo: A large language model for traditional chinese medicine](#). *Expert Systems with Applications*, 284:127672.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864. Association for Computational Linguistics.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems 36*, pages 28541–28564.
- Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yanan He, Zhangwei Gao, Erfei Cui, et al. 2025. [Omnimorphus: A unified multimodal corpus of 10 billion-level images interleaved with text](#). In *The Thirteenth International Conference on Learning Representations*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Can Liu, Kaijie Sun, Qingqing Zhou, Yuchen Duan, Jianhua Shu, Hongxing Kan, Zongyun Gu, and Jili Hu. 2024b. [Cpmi-chatglm: parameter-efficient fine-tuning chatglm with chinese patent medicine instructions](#). *Scientific Reports*, 14(1):6403.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Advances in neural information processing systems*, 36:34892–34916.
- Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. 2026. [Teaching multimodal LLMs to comprehend 12-lead electrocardiographic images](#). *npj Digital Medicine*, 9(1):349.
- Tingting Liu, Xiang Li, Zhi-Yong Zou, and Changwei Li. 2015. [The prevalence and determinants of using traditional chinese medicine among middle-aged and older chinese adults: Results from the china health and retirement longitudinal study](#). *Journal of the American Medical Directors Association*, 16(11):1002.e1–1002.e5.
- George B Moody and Roger G Mark. 2001. [The impact of the mit-bih arrhythmia database](#). *IEEE engineering in medicine and biology magazine*, 20(3):45–50.
- Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. [A time series is worth 64 words: Long-term forecasting with transformers](#). In *International Conference on Learning Representations*.
- Jorge Oliveira, Francesco Renna, Paulo Costa, Marcelo Nogueira, Ana Cristina Oliveira, Andoni Elola, Carlos Ferreira, Alipio Jorge, Ali Bahrami Rad, Matthew Reyna, Reza Sameni, Gari Clifford, and Miguel Coimbra. 2022. [The circor digiscope phonocardiogram dataset](#). *PhysioNet*. Version 1.0.3.
- Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. [The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms](#). *Scientific Data*, 8(1):156.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. [Openbiollms: Advancing open-source large language models for healthcare and life sciences](#). Hugging Face.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Alal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems 37*, pages 30811–30849.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Minghui Shi and Changle Zhou. 2007. [Diagnosis in traditional chinese medicine using artificial neural networks: State-of-the-art and perspectives](#). In *Third International Conference on Natural Computation (ICNC 2007)*, volume 2, pages 290–294. IEEE.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Siboj Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. 2025. [Gmai-v1-r1: Harnessing reinforcement learning for multimodal medical reasoning](#). *arXiv preprint arXiv:2504.01886*.

- Yang Tan, Zhixing Zhang, Mingchen Li, Fei Pan, Hao Duan, Zijie Huang, Hua Deng, Zhuohang Yu, Chen Yang, Guoyang Shen, et al. 2024. [Medchatzh: A tuning llm for traditional chinese medicine consultations](#). *Computers in Biology and Medicine*, 172:108290.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Haoyu Tian, Kuo Yang, Xin Dong, Chenxi Zhao, Mingwei Ye, Hongyan Wang, Yiming Liu, Minjie Hu, Qiang Zhu, Jian Yu, Lei Zhang, and Xuezhong Zhou. 2024. [Tcmllm-pr: evaluation of large language models for prescription recommendation in traditional chinese medicine](#). *Digital Chinese Medicine*, 7(4):343–355.
- Zhikui Tian, Dongjun Wang, Xuan Sun, Yadong Fan, Yuanyuan Guan, Naijin Zhang, Mi Zhou, Xianyu Zeng, Yin Yuan, Huaian Bu, et al. 2023. [Current status and trends of artificial intelligence research on the four traditional chinese medicine diagnostic methods: a scientometric study](#). *Annals of Translational Medicine*, 11(3):145.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. [Towards generalist biomedical ai](#). *Nejm Ai*, 1(3):AIoa2300138.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. [Hu-atuo: Tuning llama model with chinese medical knowledge](#). *Preprint*, arXiv:2304.06975.
- Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: A lightweight multilingual medical LLM towards democratizing medical AI to 6b people. *arXiv preprint arXiv:2403.03640*.
- Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, Anningzhe Gao, and Benyou Wang. 2025. [Huatu-26M, a large-scale chinese medical QA dataset](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3828–3848. Association for Computational Linguistics.
- Xu Wang, Jingwei Liu, Chaoyong Wu, Junhong Liu, Qianqian Li, Yufeng Chen, Xinrong Wang, Xinli Chen, Xiaohan Pang, Binglong Chang, et al. 2020. [Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark](#). *Computational and structural biotechnology journal*, 18:973–980.
- Sibo Wei, Xueping Peng, Yifei Wang, Tao Shen, Jisheng Si, Weiyu Zhang, Fa Zhu, Athanasios V. Vasilakos, Wenpeng Lu, Xiaoming Wu, and Yinglong Wang. 2025. [Biancang: A traditional chinese medicine large language model](#). *IEEE Journal of Biomedical and Health Informatics*, pages 1–12.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. [Pmc-llama: toward building open-source language models for medicine](#). *Journal of the American Medical Informatics Association*, 31(9):1833–1843.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. [Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data](#). *Nature Communications*, 16(1):7866.
- Jiacheng Xie, Yang Yu, Yibo Chen, Hanyao Zhang, Lening Zhao, Jiaxuan He, Lei Jiang, Xiaoting Tang, Guanghui An, and Dong Xu. 2025. [Bencao: An instruction-tuned large language model for traditional chinese medicine](#). *arXiv preprint arXiv:2510.17415*.
- Ming Xu. 2022. [text2vec: A tool for text to vector](#).
- Ming Xu. 2023. [Medicalgpt: Training medical gpt model](#). GitHub repository.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, et al. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Guoxing Yang, Xiaohong Liu, Jianyu Shi, Zan Wang, and Guangyu Wang. 2024b. [Tcm-gpt: Efficient pre-training of large language models for domain adaptation in traditional chinese medicine](#). *Computer Methods and Programs in Biomedicine Update*, 6:100158.
- Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024c. [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.
- Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang, and Yuan Luo. 2019. [Traditional chinese medicine clinical records classification with bert and domain specific corpora](#). *Journal of the American Medical Informatics Association*, 26(12):1632–1636.
- Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. [AnyGPT: Unified multimodal LLM with discrete sequence modeling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 9637–9662. Association for Computational Linguistics.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. [HuatuoGPT, towards taming language model to be a doctor](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885. Association for Computational Linguistics.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, and 4 others. 2024a. [A generalist vision-language foundation model for diverse biomedical tasks](#). *Nature Medicine*, 30(11):3129–3141.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024b. [Ultramedical: Building specialized generalists in biomedicine](#). *Advances in Neural Information Processing Systems*, 37:26045–26081.
- Changbo Zhao, Guo-Zheng Li, Chengjun Wang, and Jinling Niu. 2015. [Advances in patient classification for traditional chinese medicine: a machine learning perspective](#). *Evidence-Based Complementary and Alternative Medicine*, 2015(1):376716.
- Qi Zhao, Ran Guo, Zixuan Fan, Linlin Hu, Zhimin Hu, and Yuanli Liu. 2023. [Medical conditions and preference of traditional chinese medicine: results from the china healthcare improvement evaluation survey](#). *Patient preference and adherence*, pages 227–237.
- Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. 2025. [ECG-Chat: A large ECG-language model for cardiac disease diagnosis](#). In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.
- Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2025. [Efficiently democratizing medical llms for 50 languages via a mixture of language family experts](#). In *The Thirteenth International Conference on Learning Representations*.
- Xingzhi Zhou, Xin Dong, Chunhao Li, Yuning Bai, Yulong Xu, Ka Chun Cheung, Simon See, Xinpeng Song, Runshun Zhang, Xuezhong Zhou, and Nevin L. Zhang. 2024. [Tcm-ftp: Fine-tuning large language models for herbal prescription prediction](#). In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4092–4097. IEEE.
- Bing Zhu and Hongcai Wang. 2010. [Diagnostics of traditional Chinese medicine](#). Singing Dragon.
- Yi Zhuang, Lingkai Yu, Nan Jiang, and Yujia Ge. 2025. [Tcm-klama: Intelligent generation model for traditional chinese medicine prescriptions based on knowledge graph and large language model](#). *Computers in Biology and Medicine*, 189:109887.

A Text Public Benchmark

To further evaluate LLMs’ TCM knowledge, we test the model on publicly available benchmarks, focusing on TCM-relevant subsets of MedQA (Chinese), CMB, CMExam, and CMMLU. As shown in Table 9, ShizhenGPT achieves the best performance among open-source models under 100B parameters. The performance gap remains small, particularly in comparison with DeepSeek-R1, highlighting the importance of using real, up-to-date exam data for more meaningful evaluation.

Model	MedQA [‡]	CMB [‡]	CMExam [‡]	CMMLU [‡]
Llama-3.1-70B-Instruct	82.2	60.3	66.9	69.5
Qwen2.5-7B-Instruct	86.6	74.7	80.4	79.5
internlm2_5-20b-chat	83.3	75.1	79.7	76.7
Yi-1.5-34B-Chat	86.1	77.3	81.7	78.9
DeepSeek-R1-Distill-14B	79.3	71.7	75.2	77.4
Qwen2.5-72B-Instruct	90.8	85.1	90.5	87.2
Qwen2.5-32B-Instruct	89.9	84.8	88.7	86.9
DeepSeek-R1-Distill-32B	85.9	78.8	82.2	83.6
GPT-4o	84.0	70.7	76.7	75.0
doubao-1.5-pro-32k-250115	93.7	91.0	94.8	90.0
DeepSeek-R1	93.7	92.2	93.5	90.6
ShizhenGPT-7B	90.0	79.0	85.3	84.4
ShizhenGPT-32B	92.2	88.7	92.9	89.7

Table 9: Results on public benchmark datasets. [‡] indicates the subset of questions related to traditional Chinese medicine (TCM).

B Related Work

Multimodal LLMs for Medical-Specific Modalities Since the release of GPT-4 (Achiam et al., 2023), numerous powerful multimodal LLMs (Liu et al., 2023; Bai et al., 2025; Team et al., 2023; Chen et al., 2024e) have emerged. To strengthen their capabilities in the medical domain, researchers have adapted these models to various medical modalities. As a core multimodal task, many MLLMs have been designed to interpret image-text inputs (Awadalla et al., 2023; Li et al., 2023; Wu et al., 2025, 2024; Chen et al., 2024b; Tu et al., 2024; He et al., 2026; Su et al., 2025), supporting common medical images such as X-rays and pathology slides. Some even extend to 3D imaging like CT and MRI (Wu et al., 2025), further expanding their applicability. Beyond imaging, medical modalities such as ECG signals are also being explored, with models now able to interpret them directly (Zhao et al., 2025) or process them as visual inputs (Liu et al., 2026) to enhance understanding.

LLMs in TCM LLMs have been widely adopted in the medical domain (Xu, 2023; Wang et al., 2023; Han et al., 2023; Wu et al., 2024; Pal and

Sankarasubbu, 2024; Labrak et al., 2024; Bao et al., 2023; Zhang et al., 2024a; Chen et al., 2024b; Wang et al., 2024; Zheng et al., 2025; Christophe et al., 2024; Singhal et al., 2023). The development of LLMs for TCM has made solid progress and yielded encouraging results. Some existing models (Hua et al., 2024; Tian et al., 2024; Yang et al., 2024b; Liu et al., 2024b; Zhou et al., 2024; Xie et al., 2025) have improved performance by collecting high-quality data on TCM prescription recommendations, syndrome differentiation, and QA, then applying instruction tuning to enable diagnosis from a TCM perspective. Other studies (Yang et al., 2024c; Dai et al., 2024; Wei et al., 2025; Tan et al., 2024; Jia et al., 2025; Dong et al., 2024) have identified a lack of TCM knowledge in existing base models. Since fine-tuning alone could not fully address this, they compiled TCM resources from online forums and classical texts, incorporating this content through continued pretraining to enhance model performance. Besides that, models, like (Zhuang et al., 2025) use RAG and knowledge graph to enhance LLMs’ ability in TCM diagnosis. These efforts have laid an important foundation for the future of TCM LLMs.

C Additional Ablation Studies

C.1 Effect of Pretraining Data Scale

To assess the impact of dataset size, we performed an ablation using different amounts of pretraining data. Due to time constraints, we report results with 10% and 100% of the full data compared to a model without pretraining.

As shown in Table 10, performance improves consistently with more pretraining data, highlighting the value of large-scale pretraining.

C.2 Modality Contribution Analysis

Table 11 tests whether one jointly tuned model retains separate pulse, electronic-nose, and audio capabilities. Joint tuning preserves performance and gives small, inconsistent changes across tasks and scales. Because each task still uses one modality at a time, this experiment does not demonstrate cross-modal synergy.

C.3 Comparison with Specialized Signal Classifiers

Table 12 evaluates whether spectrogram conversion discards too much modality-specific structure. PatchTST (Nie et al., 2023) is competitive and is

Model	2024 TCM Pharmacist Exam	2024 TCM Physician Exam	TCM Visual Medicinal Recognition	TCM Visual Diagnosis
ShizhenGPT-7B				
+ 0% Pretraining Data	66.9	68.8	43.7	59.3
+ 10% Pretraining Data	68.1 (+1.2)	70.2 (+1.4)	44.0 (+0.3)	61.4 (+2.1)
+ 100% Pretraining Data	70.1 (+3.2)	76.7 (+7.9)	52.7 (+9.0)	64.5 (+5.2)

Table 10: Effect of pretraining data scale on downstream TCM tasks.

Scale	Fine-tuning Setting	Pulse Pregnancy (%)	E-nose Disease (%)	Heart Sound Abnormality (%)
7B	Modality only	79.5	47.5	63.4
7B	Joint modalities	80.1	47.9	62.9
32B	Modality only	79.2	49.2	64.6
32B	Joint modalities	80.3	50.7	65.3

Table 11: Modality ablation at both model scales. “Modality only” trains a separate model for each column. “Joint modalities” trains one model on all three.

Method	Pulse Pregnancy	E-nose Disease	ECG Abnormality	Average
Random	50.0	14.3	50.0	38.1
CNN (spectrogram)	76.2	41.0	68.7	62.0
Raw values as text to LLM	52.4	22.8	45.9	40.4
Raw signal + PatchTST	78.6	52.2	70.9	67.2
Mel + Whisper + modality token (ours)	80.5	48.8	71.5	66.9

Table 12: Comparison of shared and modality-specific signal representations under the same label spaces and data splits.

best on the electronic-nose task, confirming that raw time-series encoders remain important. The Whisper interface is comparable on average while retaining native speech support and one signal pathway. We therefore treat it as a workable shared interface, not an optimal representation for every sensor.

For each task in Table 13, we construct five stratified folds, train on four, and evaluate on the held-out fold, so every sample used for cross-validation is evaluated once. For the electronic-nose task, the folds use the 672-sample training portion of the seven-class disease dataset. Label proportions are maintained as closely as sample counts permit. The single-split values are close to the cross-validation means. The larger heart-sound variance reflects its small sample size.

D General-domain Dataset

This section introduces the general-domain pre-training data incorporated during the pre-training phase to preserve the model’s general capabilities, following (Zhang et al., 2024b; Chen et al., 2025a). Due to the extensive scale of pre-training, prolonged domain-specific learning without revis-

iting general knowledge can lead to degradation of the LLM’s foundational abilities. To address this, we included the following two categories of general-domain datasets.

General Text Dataset To maintain general language proficiency, we employed the FineWeb-edu corpus (Penedo et al., 2024), a high-quality English pre-training dataset. A randomly sampled subset comprising 5.63 billion tokens was used in mixed pre-training to support the retention of general capabilities.

General Image-Text Dataset To enhance multimodal understanding during the second stage of pre-training, we incorporated general-purpose vision-language datasets. These included ShareGPT-4V (Chen et al., 2024d), LLaVA Instruction Data (Liu et al., 2023), ALLaVA-Caption-LAION-4V (Chen et al., 2024a), and the PubMedVision Alignment Dataset (Chen et al., 2024b).

General Audio Dataset To facilitate alignment between speech and text, we introduced English general-domain audio-text pair datasets. Specif-

Task (sample size)	80/20 split	5-fold mean	Std.
Cardiac abnormality, heart sound ($n = 189$)	62.9	62.6	3.6
Disease classification, electronic nose ($n_{\text{train}} = 672$)	48.8	47.9	3.2
COVID-19 detection, cough sound ($n = 456$)	58.7	57.8	2.7
Pregnancy detection, pulse ($n = 3,165$)	80.5	78.7	1.8

Table 13: Single-split results and five-fold stratified cross-validation. Folds are patient-disjoint whenever subject identifiers are available. The in-house pulse records contain one retained record per participant.

ically, we utilized the dataset from AnyInstruct (Zhan et al., 2024).

E Pre-training Dataset

The prompt for rating TCM text quality

****Task Description:****
Please evaluate the quality of the following paragraph related to Traditional Chinese Medicine (TCM) and assign a score (1–5) based on the criteria below:

****Scoring Criteria:****
****1 point****: The content is largely irrelevant to TCM or fails to provide any useful TCM knowledge.
****2 points****: Contains minimal TCM content, or the TCM knowledge provided is extremely limited.
****3 points****: Includes some TCM content, but the information is limited or contains advertisements or false/misleading claims (unreliable).
****4 points****: Provides relatively reliable TCM knowledge and is suitable for model training.
****5 points****: Offers rich and accurate TCM knowledge of high quality, suitable for high-quality training data.

Please output a ****number between 1 and 5**** as the score only. Do not include any additional text.

****Text to be evaluated:****
 <TCM Paragraph>
 {TCM Paragraph}
 </TCM Paragraph>

Figure 4: The prompt for rating the quality of TCM documents. Here, {TCM Paragraph} represents the TCM text to be scored.

TCM Text Corpus We initially gathered approximately 5,000 TCM books and used MinerU to convert scanned PDFs into interleaved text–image corpora. After OCR and quality filtering, 3,256 books were retained. Removing covers, tables of contents, and other non-informative content produced the final 3.8GB Book Corpus. For online data, we constructed a 30K-term TCM lexicon to identify high-density TCM documents from Common Crawl (2017–2023) and WeChat public articles (primarily sourced from OmniCorpus-CW (Li et al., 2025)), yielding 33.2GB and 63.2GB, respectively. We then (1) removed low-quality or spammy doc-

uments using a classifier trained on DeepSeek-V3 labels with text2vec-base-chinese (Xu, 2022) and the rubric in Figure 4, and (2) performed embedding-based semantic deduplication. The final TCM web corpus contains 21.2GB (5B tokens).

TCM Image-Text Corpus We extract 17.6GB of interleaved image-text data (51K images) from TCM books with rich illustrations (>50 images per book). Additionally, from 5 million WeChat articles containing images, we filtered TCM-relevant samples using an image classifier based on a CLIP vision encoder fine-tuned with GPT-4o-labeled data. This process yielded 140.7GB of web-based image-text data (1M images). To enhance alignment quality, we followed HuatuoGPT-Vision and used Doubao-1.5-vision-pro (a multimodal LLM) to synthesize descriptive captions based on image and context using the prompt in Figure 5, generating 40.6GB of synthetic high-quality image-text pairs.

The prompt for generating descriptive captions for TCM images

{image}

Please generate a professional, detailed, and high-quality description for the medical image I provide. The description should include as many Traditional Chinese Medicine (TCM)-related visual details as possible to ensure clinical readability, professionalism, and comprehensiveness, while also being understandable to the general public. Aim to make the description as rich and detailed as possible, providing extensive visual information.

You may refer to the image’s context to improve the accuracy and completeness of your description, but do not reveal that you referenced the context.

<caption>
 {caption}
 </caption>

<image context> (<image> represents the location of the image)
 {image context}
 </image context>

Please output a detailed description of the image only. Do not generate any content unrelated to the task.

Figure 5: The prompt for generating descriptive captions for TCM images. Here, `image` represents the TCM image. `caption` and `image context` represent the corresponding image caption and contextual information.

TCM Audio-Text Corpus We extracted 60K TCM-related doctor-patient dialogues from the Huatuo-26M (Wang et al., 2025) dataset and synthesized corresponding audio using a high-fidelity TTS system (Du et al., 2024). This produced 58K audio-text pairs aligned at the utterance level, forming a TCM-specific audio-text dataset.

F Pre-Training Details

Stage 1: Text-Only Pre-training Since textual data greatly outweighs multimodal data in token volume, we begin by pretraining the model on TCM-related textual corpora to establish strong domain knowledge. This stage uses 6.29B tokens from TCM sources, including web and books, supplemented with 5.63B tokens of general-purpose text to preserve the LLMs’ foundational capabilities. As detailed in Appendix D, incorporating general data does not impair TCM-specific learning, while effectively preserving the model’s general reasoning abilities. In total, Stage 1 involves 11.92B tokens of text-only pretraining. To optimize training efficiency, we concatenate input samples into sequences of 4096 tokens, using the special token `<|endoftext|>` to delimit different documents. The model is trained to predict all tokens except the `<|endoftext|>` tokens. Pretraining is conducted for 1 epoch with a batch size of 256, a learning rate of $5e-5$, and a warmup ratio of 0.005.

Stage 2: Multimodal Pre-training We then pre-train on multimodal data to introduce TCM visual knowledge and basic audio understanding. This includes 1.17B tokens from TCM image-text data and 0.69B tokens from general image-text datasets (e.g., ShareGPT-4V), totaling 1.86B image-text tokens. For audio-text alignment, we include 0.03B tokens from TCM-specific audio-text data and 0.04B from general audio datasets to ensure basic audio comprehension. Additionally, we resample 1.8B tokens from the Stage 1 text corpus to maintain textual grounding during multimodal learning. Altogether, Stage 2 involves 3.84B tokens of multimodal pre-training. As in Stage 1, inputs are packed into

4096-token sequences to accelerate training. And only textual tokens are used for loss computation. We use a batch size of 128 and the same learning rate of $5e-5$ for this stage.

G Data Leakage Prevention

Preventing data leakage between training and evaluation is critical for reliable benchmarking. In this work, we adopted two strategies to mitigate leakage risk:

G.0.1 Source and Temporal Separation

The major text pre-training sources were collected no later than 2023.

- **Common Crawl:** 2017–2023 (no data from 2024 or later).
- **WeChat articles:** collected prior to June 2023.
- **TCM textbooks:** published no later than 2021.

Evaluation benchmarks were assembled separately from recent official exam papers and TCM atlases. The eight atlases used to construct the TCM-Vision Benchmark were explicitly excluded from training. Because an exam’s nominal year and administration date may differ, we do not use the year label alone as evidence of temporal separation.

G.0.2 Content-Based Filtering

To further prevent leakage, we applied a strict content overlap filter. For the TCM-Vision Benchmark, we followed the protocol of Med-PaLM 2 (Singhal et al., 2025), removing any evaluation instance with a 64-character or longer exact match with any training source, including textbooks and scraped corpora.

These source-separation and content-overlap checks reduce the risk of benchmark leakage, but they do not rule out all semantic overlap between general training corpora and evaluation questions.

H Instruction Fine-tuning Dataset

The Prompt for Verifier

```
<Model Response>
{Model Response}
</Model Response>

<Reference Answer>
{Ground-truth Answer}
</Reference Answer>
```

You are provided with a model-generated response (<Model Response>) and a reference answer (<Reference Answer>). Compare the model response with the reference answer and determine its correctness. Your task is to simply output "True" if the response is correct, and "False" otherwise.

Figure 6: The prompt for the DeepSeek-V3 verifier. {Model Response} represents the output of the model to be verified. {Ground-truth Answer} represents the ground-truth answer or doctor response for TCM problems.

Text Instructions We collected 92k authentic TCM-related questions from the Huatuo-26M doctor-patient dialogue corpus and the HuatuoGPT-o1 verifiable question set. These questions were then distilled using DeepSeek-R1 to extract formal responses, omitting the reasoning components. Given that each question was paired with either a doctor’s reply or a verifiable answer, we employed DeepSeek-V3 to assess whether the responses passed verification, using the prompt illustrated in Figure 6. This process ultimately yielded a Chinese instructional dataset comprising 83k entries.

The prompt for generating questions for TCM Vision Instructions

{image}

Please generate a Traditional Chinese medicine (TCM)-related question about the medical image I provide. The question should assess the model’s visual capabilities. Avoid being too specific—design the question so that it requires the model to look at the image to answer. The question should demand strong visual understanding as well as some knowledge of Traditional Chinese medicine.

You may refer to the provided image caption and contextual information to improve the quality of the question. However, ****do not mention or reference the caption or context in the question itself—assume they are not available****.

<caption>
{caption}
</caption>

<image context> (<image> represents the location of the image)
{image context}
</image context>

Please generate the question directly. Do not include any content unrelated to the task.

Figure 7: The prompt for generating questions for TCM Vision Instructions. Here, image represents the TCM image. caption and image context represent the corresponding image caption and contextual information.

Vision Instructions We filtered TCM-related images with detailed captions from the books and web interleaved data. GPT-4o first generated vision-related questions grounded in each image, caption, and context using the prompt in Figure 7, then generated the corresponding answer using Figure 8. This process yielded 65K visual instruction samples.

The prompt for generating answers for TCM Vision Instructions

{image}

<question>
{question}
</question>

You are now required to look at the image I provide and answer the user’s question about Traditional Chinese Medicine (<question>). Make sure your response directly addresses the user’s query, follows instructions well, and is as detailed and rich as possible, with the style and quality characteristic of GPT-4o.

You may refer to the image caption and contextual information I secretly provide to you in order to improve the accuracy and completeness of your answer. However, ****do not mention or reference the caption or context in your response—assume they are not available****.

<caption>
{caption}
</caption>

<image context> (<image> represents the location of the image)
{image context}
</image context>

Please generate the answer directly. Do not include any unrelated content.

Figure 8: The prompt for generating answers for TCM Vision Instructions. Here, image represents the TCM image. question represents the generated visual question. caption and image context represent the corresponding image caption and contextual information.

H.1 Synthetic Visual Supervision Verification

To assess grounding fidelity, we used GPT-5.4 (high) to check a stratified sample of 1,000 synthesized image–text items against the source image and atlas or web reference. Table 14 reports the results. This is an LLM-based spot check rather than verification by TCM practitioners. It therefore measures reference consistency rather than medical correctness. The vision instruction set (about 71K images) is also much smaller than the approximately 1.2M TCM images used for multimodal

pretraining, so its primary role is instruction alignment.

Synthesized data	Checked	Grounding-consistent
Vision QA (GPT-4o)	500	96.6%
Image captions (Doubao)	500	93.6%

Table 14: Reference-grounded spot check of synthetic visual supervision.

Reference-grounded verification prompt

```
{image}
<reference> {caption / context} </reference>
<synthesized> {generated QA or caption} </synthesized>
Judge whether the synthesized content is faithful to the image and reference: it must not introduce claims that contradict or are unsupported by the evidence. Output only "Faithful" or "Unfaithful", followed by a one-line reason.
```

Audio Instructions The audio instruction set includes two parts. For speech instruction, we sampled 20K questions from the text instructions and converted them into speech using TTS. Each audio was paired with its original answer. Additionally, we added transcription-style prompts (e.g., "Please repeat exactly what was said") to the same audio inputs, creating another 20K instruction pairs. For TCM-specific audio, such as cough COUGHVID or heart sounds CirCor DigiScope, we converted these classification tasks into instruction format. For example: "Please listen to the patient's cough and determine if they have COVID-19. Answer: Yes or No."

Physiological Signal Instructions For non-audio physiological signals, we convert all inputs into waveform-like sequences and prepend modality-specific tokens (e.g., "<Smell>", "<Pulse>", "<ECG>"). For electronic-nose input, we use the authorized CUHKSZ-Odors dataset. The in-house pregnancy-pulse dataset contains 3,165 participants and uses an 80/20 split. The label comes from obstetric records rather than a subjective pulse judgment. ECG data come from the PTB Diagnostic ECG Database and MIT-BIH Arrhythmia Database. All classification tasks are reformulated as instructions. These tasks evaluate representation learning and must not be interpreted as clinical diagnostic tests.

I Instruction Fine-tuning Details

To complete the training pipeline, we merge all modality-specific instruction datasets (Table 3) and

fine-tune the TCM-pretrained model jointly. All instructions are formatted using the Qwen2.5 prompt template, and only the output tokens corresponding to the model's response are included in the loss computation. The model is trained for 3 epochs with a learning rate of 5e-6, batch size of 128, and a warmup ratio of 4%. This stage yields our final model, **ShizhenGPT**.

J Physiological Signal Datasets

Pulse Pulse signals capture fluctuations in cardiac rhythm and have historically played a vital role in assisting diagnosis within the framework of TCM.

- **Pregnancy Identification Dataset:** In collaboration with a partner hospital, pulse signals were collected from 3,165 women of child-bearing age by licensed TCM practitioners with over two years of clinical experience. Data collection was conducted at rest using a pulse acquisition device at the Cun, Guan, and Chi positions on both radial arteries. Each retained record corresponds to one participant. Pregnancy and gestational-age labels come from obstetric records, with gestational age determined by the last menstrual period or ultrasound. The 80/20 split is therefore patient-disjoint. This dataset evaluates pulse-signal representation and is neither external clinical validation nor a replacement for hCG testing or ultrasound.
- **AdBrc pulse** (Guo et al., 2022) dataset: We adopted this publicly available dataset to enrich the pulse diagnosis task. It includes nine health conditions, each with 218 samples, with signals collected from the Guan position based on the TCM Cun-Guan-Chi theory. This version provides two types of pulse signals: continuous and cycled. We used only the continuous signals from the "continues" folder, which contain raw data acquired via pressure sensors. All data were converted into a unified time-series format and split into training and testing sets using an 80/20 ratio.

Smell We used the **CUHKSZ-Odors** dataset to evaluate the model's ability to interpret olfactory signals in TCM. The dataset, intended solely for research and requiring prior authorization via the official platform, must not be used commercially or shared with unauthorized parties. The original

dataset contains eight categories: Health, Diabetes, Nephrosis, Bile Diseases, Liver Disease, Lung Disease, Gastritis, and Digestive Diseases. Each category contains 120 samples, including 100 clean samples and 20 samples with sampling interference. For our seven-class disease-classification experiment, we exclude the Health category and retain 840 samples from the seven disease categories. The 80/20 split yields 672 training instructions and 168 test samples, which is consistent with the 14.3% random baseline. Each sample is a 1153×9 matrix collected by an electronic nose sensor array, where the nine channels correspond to gas sensors that record temporal responses to volatile organic compounds. Therefore, the smell input is not a text description, but a multi-channel time-series signal. In our pipeline, this matrix is converted into the unified waveform representation and tagged with the modality-specific <Smell> token before being processed by the signal encoder. In TCM, smelling diagnosis focuses on body odor and breath-related cues. The electronic nose provides a quantified proxy by capturing reproducible chemical response patterns associated with these olfactory signals.

Sound In TCM auscultation, patient voice characteristics are an important diagnostic cue. To assess our model’s audio classification ability, we used the **COUGHVID dataset** (Orlandic et al., 2021), which contains over 25,000 crowdsourced cough recordings across diverse demographics and COVID-19 statuses. Of these, 2,800 samples were expert-labeled by physicians, forming one of the largest annotated cough datasets for respiratory condition detection. This resource supports robust model evaluation in sound-based diagnostic tasks. To address the severe class imbalance, we adjusted the dataset to achieve equal numbers of samples for both labels. An 80/20 split was then applied for training and testing.

Modern Medicine Modalities Given the limited scale of TCM-specific signals, we add modern biomedical time series to test whether the interface extends beyond the training modality. These are not presented as traditional TCM modalities.

- **Electrocardiogram (ECG) signals** are modern biomedical measurements of cardiac rhythm. We use the **PTB Diagnostic ECG Database** (Bousseljot et al., 1995) and **MIT-BIH Arrhythmia Database** (Moody and Mark, 2001) only to test extensibility to

analogous time-series inputs, using their original train/test splits where provided.

- **Heart sounds** are evaluated with **The CirCor DigiScope Phonocardiogram Dataset** (Oliveira et al., 2022). This is a preliminary listening-related classification task, not a complete model of TCM auscultation. We use an 80/20 split and additionally report five-fold results in Table 13.

K Benchmarks

To comprehensively evaluate the TCM capabilities of our model, we design a comprehensive multi-modal benchmark suite across all modalities: text, vision and physiological signals. The benchmarks are categorized as follows:

- **Text Benchmarks** We evaluate textual TCM knowledge using two sources: (1) official national exams and (2) public QA benchmarks. For the former, we collect recent national exams in China, including the 2024 National TCM Pharmacist Licensing Exam, the 2024 TCM Physician Licensing Exam, and the 2024 Assistant Medical Exam. These exams were administered after our training data collection cutoff to ensure temporal separation and minimize training overlap.
- **Vision Benchmark** To assess visual understanding in TCM, we construct a benchmark from eight authoritative atlases explicitly excluded from training. Each book contributes to a specific task: *Chinese Herbal Medicine Color Atlas* (herb recognition), *Chinese Medicinal Materials* (TCM material classification), *Practical Atlas of Decoction Pieces* (decoction recognition), *Holographic Inspection Atlas* (holistic diagnosis), *Tongue Diagnosis Atlas* (tongue inspection), *Eye Diagnosis Atlas* (eye pattern recognition), *Palm Symptom Atlas* (palm inspection), and *Tuina Techniques Atlas* (massage gesture recognition). All visual questions use a multiple-choice format: one image, one question, four options, and one correct answer. For images with explicit captions (e.g., herbs), we randomly select three nearby captions as distractors and use the source caption as the answer. For more complex content (e.g., Tuina), GPT-4o generates a question grounded in the image, atlas caption, and local context. The final

benchmark contains 7,204 questions across eight TCM-specific visual domains. Table 15 reports construction routes, and Table 16 reports a manual check against atlas evidence.

Construction route	Count	Fraction
Caption-derived recognition	3,239	45.0%
GPT-4o visual reasoning	3,965	55.0%

Table 15: TCM-Vision benchmark construction. Answers in both routes are grounded in atlas captions. Subset labels will be released.

Manual check on 300 sampled items	Count
Fully supported by image + atlas evidence	283 (94.3%)
Not fully supported	12 (4.0%)
Ambiguous options	5 (1.7%)

Table 16: Manual answerability check against atlas annotations.

- **Human Benchmark** Details at Appendix L.
- **Signal Benchmark** For signal-based modalities such as smell, auscultation, and ECG, we evaluate on the following datasets: COUGHVID (Orlandic et al., 2021) for respiratory sounds and cough, The CirCor DigiScope Phonocardiogram Dataset (Oliveira et al., 2022) for heart sounds, CUHKSZ-Odors for smell signals, PTB Diagnostic ECG Database (Bousseljot et al., 1995) and MIT-BIH Arrhythmia Database (Moody and Mark, 2001) for ECG signals. We use official test splits when provided. Otherwise, we create patient-disjoint splits when subject identifiers are available and report five-fold stratified cross-validation for the four smaller or custom-split tasks in Table 13. For the in-house pregnancy-pulse task, 20% of participants are held out for testing.

L Human Evaluation

We selected 6 powerful LLMs to compare against ShizhenGPT, using a set of 90 real-world TCM cases from *Journal of Chinese Clinical Case Achievements Database*. ShizhenGPT’s responses were compared pairwise with those of each model, resulting in 540 pairwise comparisons. After randomly shuffling the model order, we invited three licensed TCM practitioners to evaluate the responses based on correctness, completeness, and clinical insight. The 540 pairwise comparisons were evenly

distributed across three doctors, with each doctor evaluating 180 pairs and receiving a compensation of 300 RMB. For each pair, they were asked to choose the response that better aligns with TCM medical standards, or mark them as equally aligned if no clear preference could be made (see Figure 9 for evaluation criteria).

The prompt for human evaluation

Please evaluate the responses of two language models to the same TCM-related question and indicate which answer you professionally prefer. Your judgment should be based primarily on the accuracy of TCM theory, the proper use of professional terminology, the completeness of the response, and its clinical relevance.

If one response is clearly more rigorous, reasonable, or clinically valuable—or if the other contains obvious errors, unclear expression, or lacks professionalism—please label it as “Model 1” or “Model 2” accordingly. If both responses are of similar quality and no clear preference can be made, please mark it as “Tie”.

During the evaluation, please avoid being influenced by writing style, expression, or response length. Focus solely on the accuracy and practical value of the TCM content. We sincerely appreciate your professional judgment and support!

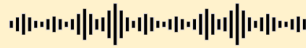
Figure 9: The instruction given to the doctors for human evaluation.

M Evaluation Protocol

For both text and vision evaluation, we run inference with seeds 0–4 and report the mean. For vision, ShizhenGPT-7B and ShizhenGPT-32B obtain 60.2 ± 0.74 and 63.3 ± 0.91 , respectively. Per-domain vision values in Table 5 show seed 0 for compactness. Training uses seed 42. Signal robustness is evaluated separately with the stratified folds described in Table 13.

N Output Examples

Figures 10 and 11 show separate audio and image examples. They are qualitative illustrations only and should not be interpreted as evidence of clinical safety, treatment efficacy, or patient-level cross-modal fusion.



I'm a middle-aged man, and lately, I've been feeling short of breath whenever I do any physical activity. At night, I often wake up because I can't breathe properly, and lying flat makes it even worse. The doctor examined me and found that my heart is beating fast, the size of my heart seems to have changed, there's a heart murmur, some abnormalities in my lungs, and swelling in my legs. In this situation, what would be the best way to manage or treat my condition?

In Traditional Chinese Medicine (TCM), your symptoms suggest a complex pattern involving multiple organ systems, primarily the Heart, Spleen, Kidneys, and Lungs. Here's a breakdown of the TCM perspective and management approach:



TCM Diagnosis (Pattern Differentiation):

1. Heart Qi/Yang Deficiency:

- * Rapid heartbeat, heart murmur, and fatigue during activity indicate weakened Heart Qi/Yang, impairing circulation and leading to blood stasis.
- * Enlarged heart may correlate with "Phlegm-Blood Stasis" obstructing the Heart vessels.

2. Spleen Yang Deficiency with Damp Accumulation:

- * Leg swelling (edema) reflects Spleen's inability to transform fluids, causing dampness and water retention.

3. Kidney Yang Deficiency:

- * Nighttime breathlessness and worsening symptoms when lying flat suggest Kidney Yang failing to "grasp Qi" and regulate water metabolism, leading to upward rebellion of fluids (e.g., pulmonary congestion).

4. Lung Qi Deficiency with Phlegm-Damp:

- * Shortness of breath and lung abnormalities relate to Phlegm-Damp obstructing the Lungs, impairing Qi diffusion.

Treatment Principles:

- * **Tonify Heart and Kidney Yang** to strengthen circulation and water metabolism.
- * **Strengthen Spleen Yang** to resolve dampness and edema.
- * **Transform Phlegm, Activate Blood**, and regulate Lung Qi to alleviate breathlessness.

Herbal Formulas (Customized Combinations):

1. Zhen Wu Tang (True Warrior Decoction):

- * Warms Kidney Yang, transforms water retention, and reduces edema. Key herbs: Fu Zi (aconite), Bai Zhu (atractylodes).

2. Ling Gui Zhu Gan Tang (Poria-Cinnamon Decoction):

- * Strengthens Spleen Yang, resolves damp-phlegm. Key herbs: Fu Ling (poria), Gui Zhi (cinnamon).

3. Modifications for Blood Stasis/Phlegm:

- * Add **Tao Hong Si Wu Tang** (peach kernel, safflower) to invigorate blood.
- * Include **Ting Li Zi** (Semen Lepidii) to drain Lung fluid and relieve breathlessness.

4. Supporting Additions:

- * **Wu Wei Zi** (Schisandra) to astringe Lung-Kidney Qi and stabilize breathing.

Acupuncture Points:

- * **Tonify Yang/Qi:** BL23 (Shenshu, Kidney), CV4 (Guanyuan), ST36 (Zusanli).
- * **Resolve Edema:** SP9 (Yinlingquan), SP6 (Sanyinjiao), CV9 (Shuifen).
- * **Regulate Heart/Lung:** PC6 (Neiguan), LU7 (Lieque), BL13 (Feishu).
- * **Moxibustion:** Apply on CV8 (Shenque) or ST36 to warm Yang and dispel cold-damp.

Lifestyle/Dietary Advice:

- * **Avoid Cold/Raw Foods:** These impair Spleen Yang. Opt for warm, cooked meals (e.g., ginger, cinnamon, pumpkin).
- * **Limit Salt and Fluids:** Reduce dampness accumulation.
- * **Gentle Exercise:** Tai Chi or Qi Gong to improve Qi circulation without overexertion.
- * **Elevate Legs:** To assist fluid drainage.

By addressing the root (Yang deficiency) and branch (phlegm, blood stasis, edema), TCM aims to restore balance and improve your quality of life. Consistency in treatment and lifestyle adjustments is key.

Figure 10: An example of ShizhenGPT with audio input



I'm 38 years old, and lately, I've been feeling a deep, lingering heat in my body, especially in the evenings. I sweat a lot at night, even when I'm asleep, and I've been experiencing involuntary ejaculation during sleep. My legs and knees often feel sore and warm, which makes it uncomfortable to walk. I've also been coughing, and sometimes I even cough up blood. On top of that, I often feel restless and irritable, and I lose my temper easily. Please check my tongue coating to get more information.



From the perspective of Traditional Chinese Medicine (TCM), the tongue appears slightly red with a thin and somewhat dry coating, along with a prominent central fissure. These features suggest a pattern of yin deficiency, possibly with internal heat. The dry tongue and fissuring indicate a lack of body fluids, often associated with prolonged illness or constitutional weakness. The absence of scalloped edges suggests that spleen qi deficiency is not prominent. Overall, this tongue presentation aligns with a diagnosis of yin deficiency with internal heat or kidney yin deficiency, which is often seen in individuals experiencing symptoms such as tidal fever, night sweats, spontaneous ejaculation, dry cough with blood, restlessness, and irritability.

Diagnosis:

Kidney Yin Deficiency with Internal Heat affecting the Lungs, leading to symptoms of deficient fire flaring upward and fluid depletion.

Treatment Principle:

1. **Nourish Kidney Yin** to anchor the root deficiency.
2. **Clear Deficient Heat** to alleviate internal fire.
3. **Moisten the Lungs** to resolve dryness and stop bleeding.
4. **Calm the Spirit** to address restlessness and irritability.

Herbal Formula:

1. Modified Zhi Bai Di Huang Wan

- * Nourishes Kidney Yin: Shu Di Huang, Shan Zhu Yu, Shan Yao.
- * Clears Heat: Zhi Mu, Huang Bai.

2. Additions for Lung Yin and Bleeding:

- * Mai Men Dong and Tian Men Dong: Moistens Lung Yin.
- * Bai He: Nourishes Lung and calms the spirit.
- * Bai Ji: Stops bleeding and heals Lung tissue.
- * E Jiao: Nourishes blood and Yin, stops bleeding.

3. Dosage:

- * Decoction or prepared pills, tailored by a licensed TCM practitioner.

Acupuncture Points:

1. **Kidney 3 (Taixi)**: Tonifies Kidney Yin.
2. **Spleen 6 (Sanyinjiao)**: Nourishes Yin of Kidneys, Liver, and Spleen.
3. **Lung 7 (Lieque)**: Benefits Lung Yin and relieves cough.
4. **Large Intestine 11 (Quchi)**: Clears heat.
6. **Heart 7 (Shenmen)**: Calms the mind and reduces irritability.

Frequency: 2–3 sessions weekly for 4–6 weeks.

Lifestyle and Dietary Advice:

1. **Avoid:** Spicy, fried, or drying foods; alcohol; excessive sexual activity; overwork.
2. **Consume Yin-Nourishing Foods:**
 - * Pear, lotus root, white fungus, tofu, black sesame, mulberries.
3. **Rest and Hydration:** Prioritize sleep, moderate exercise (e.g., tai chi), and stay hydrated with room-temperature water or herbal teas (e.g., chrysanthemum or lily bulb tea).
4. **Mindfulness Practices:** Meditation or deep breathing to manage stress and irritability.

Explanation of Pattern:

- * **Kidney Yin Deficiency:** The root cause, leading to internal heat (night sweats, warm limbs, nocturnal emissions).
- * **Lung Involvement:** Kidney Yin fails to nourish the Lungs, causing dry cough and blood-streaked phlegm.
- * **Tongue Signs:** Red tongue with dry coating and fissures confirms Yin deficiency and fluid loss.

Follow-Up:

- * Monitor reduction in night sweats, cough, and irritability.
- * Adjust herbs if heat symptoms persist (e.g., add Di Gu Pi for deeper heat clearance).
- * Address chronic habits (e.g., stress management) to prevent relapse.

Note: If blood in cough persists, recommend Western medical evaluation to rule out tuberculosis or other pathologies.

This holistic approach aims to restore Yin-Heat balance, strengthen the Kidneys and Lungs, and improve overall vitality.

Figure 11: An example of ShizhenGPT with image input