

# TCMQA: A 38K-Question Traditional Chinese Medicine Benchmark with a Licensed-Practitioner Reference

Tzu-Heng Huang<sup>†,\*</sup>    Jet Lin<sup>‡,\*</sup>    Eric Lin<sup>§</sup>

<sup>†</sup>University of Wisconsin-Madison

<sup>‡</sup>University of California, Merced

<sup>§</sup>TechTCM

September 29, 2026

## Abstract

Medical benchmarks for language models are built almost entirely on Western biomedicine. Traditional Chinese Medicine (TCM) is a separate system, with its own diagnostic framework and its own literature, and it remains largely unmeasured. The few TCM evaluations that exist are small, narrow, and rarely paired with a human reference. We present TCMQA, an open benchmark of 38,279 questions from Chinese TCM licensing examinations, paired with 15,151 responses from 101 licensed practitioners. We evaluate 29 instruction-tuned models from 9 families, spanning 0.27B to 14.8B parameters. Accuracy ranges over 59 points, and no model approaches saturation. Pretraining data predicts TCM ability far better than scale: a 12B Western-pretrained model reaches 39.6%, while a Chinese-pretrained model an eighth its size reaches 60.8%. Nine models exceed the practitioner majority vote of 64.9%, the best by 21.8 points, and all nine come from that same Chinese-pretrained family. Yet difficulty does not transfer between models and practitioners: accuracy is flat across practitioner-rated difficulty, item-level agreement is near zero for all 29 models, and on 8.4% of items the practitioners are correct where the leading model is wrong. We release the corpus, the practitioner responses, the harness, and per-item model outputs at <https://huggingface.co/datasets/TechTCM/TCMQA>.

## 1 Introduction

Large language models (LLMs) now perform competitively on open-ended question answering, complex reasoning, and specialized tasks in physics, law, and medicine. These models are now part of daily medical use: patients ask them for diagnostic suggestions and help reading prescriptions, and clinicians use them for case analysis and documentation [1, 2, 3, 4, 5, 6, 7]. A large body of medical benchmarks has followed, evaluating factual knowledge, safety, and clinical reasoning [8, 9, 10, 11]. Most of this work assumes Western biomedicine, whether the questions are posed in English or translated into other languages. Traditional Chinese Medicine (TCM), a separate medical system with its own diagnostic framework and its own therapeutic principles, remains largely unmeasured.

Normal users bear most of the cost of this gap. Patients now ask general-purpose assistants which herb suits a presenting pattern, whether two formulas conflict, and what a prescription written by their clinician means [12, 13, 14]. They ask in Chinese, and the models answer fluently in Chinese whether or not they know any TCM. Fluency, however, says nothing about whether an answer is grounded in TCM, and few leaderboards report TCM ability, so users have no practical way to check. Privacy concerns push many toward open-weight models they can run locally—often the smaller models least likely to have absorbed TCM knowledge during pretraining.

Existing TCM evaluations do not close this gap either [15, 16]. They are typically small—a few thousand items—narrow in coverage, and either closed or only partly released. Each of these limits what a score can tell us: a small test set cannot separate models that finish close together, thin coverage leaves whole areas of the field untested, and a closed set cannot be reproduced. Nor do these benchmarks report a human reference point. Accuracy is hard to interpret in isolation: the natural comparison is what licensed practitioners score on the same items.

We introduce TCMQA, an open benchmark built from Chinese TCM licensing examinations, with three components. The *corpus* holds 38,279 multiple-choice items in a uniform five-option format, covering the professional curriculum from materia medica and formula composition through classical texts and professional regulation; 2,386 of

\*Equal contribution. Contact through: [thuang273@wisc.edu](mailto:thuang273@wisc.edu) & [jetlin101@gmail.com](mailto:jetlin101@gmail.com)

them have more than one correct option. The *human reference* comes from 101 licensed practitioners, who contributed 15,151 responses through an annotation platform we built. On 4,801 items three practitioners responded independently, and those items form the subset for every human–model comparison. Practitioners also supplied domain labels and a difficulty rating recorded before the answer was scored. The *harness* scores each model in a single pass, reading a prediction both from the generated text and from the first-token log-probabilities. We apply it to 29 instruction-tuned models from 9 families, ranging from 0.27B to 14.8B parameters and spanning Western- and Chinese-developed systems.

The benchmark separates the field by 59 accuracy points, from 18.95% to 77.98%, and no model saturates it. We find that pretraining composition matters, not scale: within a family accuracy rises monotonically with size, but across families size predicts little. Gemma-3-12B-It [17] reaches 39.60%, while Qwen2-1.5B-Inst. [18], with an eighth of the parameters, reaches 60.81%; the OLMo family [19, 20] never clears 27% at any size we tested. Nine models exceed the practitioner majority vote of 64.94%, the strongest by 21.79 points, and all nine descend from a single Chinese-pretrained family.

Moreover, practitioners agree substantially about which items are hard, yet model accuracy is flat across that consensus: on the 616 items all three practitioners missed, the model leading the subset scores 87.34%, slightly above its 86.73% average on the subset as a whole. Their difficulty ratings say the same thing: practitioner accuracy falls 57 points from the easiest level to the hardest, while most model lines barely move. Models are furthest ahead on Regulation area, which rewards recall, and closest on diagnosis and pattern differentiation, which needs clinical reasoning.

We make the following contributions.

- **An open TCM benchmark at scale.** 38,279 items in a uniform five-option format, single- and multi-answer, released in full—an order of magnitude larger than prior TCM evaluations.
- **A licensed-practitioner reference.** 15,151 responses from 101 practitioners, 4,801 items with three independent annotations each, plus domain labels, difficulty ratings, and quality flags—to our knowledge the first for TCM.
- **A reproducible testbed.** A fixed single-pass protocol that scores each model from both its generated text and its log-probabilities, applied to 29 instruction-tuned models from 9 families.

## 2 Dataset and Evaluation Testbed

TCMQA comprises three components: (i) a corpus of 38,279 examination items, (ii) a human reference collected from licensed practitioners through an in-house annotation platform, and (iii) an automated harness that evaluates models under a fixed protocol. We describe each in turn.

### 2.1 Dataset Construction

Our corpus is assembled from two sources of Chinese TCM licensing examination material: an in-house collection of practice examinations used to prepare candidates for licensure, and a web-crawled archive of historical licensing papers. Much of this material exists only in PDF form. We process it with QWEN2.5-14B-INSTRUCT, prompting the model to read each document and convert each item into a JSON record under a fixed schema: the question, five labeled options, the reference answer, and any explanation or difficulty value the source supplied. Extracted records then pass through a schema filter: we keep an item only if it has exactly five options and a reference answer. Our human study also serves as an independent audit, as practitioners could flag any item as defective or incorrect, and we exclude flagged items from the final results.

Each retained item is a tuple  $(q, \mathcal{C}, y)$ , where  $q$  is the question,  $\mathcal{C} = \{c_A, \dots, c_E\}$  is the option set, and  $y \subseteq \Sigma$  is the ground-truth answer set over the option alphabet  $\Sigma = \{A, B, C, D, E\}$ . Every item has exactly five options, so the random-guess floor is uniform across the benchmark. We include two types of question: the majority have a single correct option (35,893 items, 93.8%), and the remainder have two or more (2,386 items, 6.2%). The multi-answer items are there to raise the difficulty of the testbed. Table 2 summarizes the composition.

### 2.2 Human Reference

An accuracy figure on a knowledge benchmark is hard to interpret without a reference point. Whether 85% on licensing material counts as competence or failure can be judged only against the people who are licensed to act on

Table 1: The ten-domain TCM taxonomy, with the domain portfolio of the 4,801-item comparison subset. *Items* is the number of subset items assigned the domain by at least one of their three raters, and *Share* is that count as a fraction of 4,801.

| #  | Domain                            | Items | Share | Scope                                                               |
|----|-----------------------------------|-------|-------|---------------------------------------------------------------------|
| 1  | Foundational Theory<br>中醫基礎理論     | 1,304 | 27.2% | Yin–Yang, Five Phases, visceral manifestation, qi and blood         |
| 2  | Classical Texts<br>中醫經典文獻         | 700   | 14.6% | The canonical medical literature and its interpretive tradition     |
| 3  | Diagnosis & Patterns<br>診斷與辨證學    | 1,309 | 27.3% | The four examinations; pattern differentiation and treatment        |
| 4  | Meridians & Acupuncture<br>經絡與針灸學 | 502   | 10.5% | Channel theory, point location, needling and moxibustion            |
| 5  | Materia Medica<br>中藥本草學           | 1,494 | 31.1% | Herb properties, formula composition, preparation, contraindication |
| 6  | Internal Medicine<br>中醫內科學        | 1,982 | 41.3% | Internal disorders, incl. gynecology, pediatrics, mental illness    |
| 7  | External Medicine<br>中醫外科學        | 698   | 14.5% | Surface disorders, dermatology, traumatology                        |
| 8  | ENT & Ophthalmology<br>中醫五官科學     | 84    | 1.7%  | Disorders of the eye, ear, nose, and throat                         |
| 9  | Regulation<br>中醫法規                | 495   | 10.3% | Licensure, practice scope, drug control, professional ethics        |
| 10 | Other<br>其他                       | 1,419 | 29.6% | Items not covered above                                             |

Table 2: Composition of the TCMQA corpus.

| Property                         | Count  | Share  |
|----------------------------------|--------|--------|
| Total items                      | 38,279 | 100.0% |
| single-answer                    | 35,893 | 93.8%  |
| multi-answer                     | 2,386  | 6.2%   |
| Items with five options          | 38,279 | 100.0% |
| Items with reference explanation | 23,582 | 61.6%  |
| Items with source difficulty     | 38,217 | 99.8%  |

the same knowledge. We built the human reference into the testbed and collected it under conditions that match the model evaluation.

We first built a web application to conduct the study. It authenticates each practitioner, assigns a fixed private block of items, presents one item at a time, and records every interaction. Items were drawn as a stratified 6,000-item subset of the corpus and partitioned into 40 blocks of 150. Practitioner  $u$  receives block  $\lfloor (u - 1)/3 \rfloor$ , so each block goes to exactly three practitioners and each practitioner answers 150 items. Several screenshots from our annotation interface are provided in Appendix A.

Each item is presented on its own screen, and a single submission records the answer together with all annotations, then advances to the next unanswered item. Beyond the answer, the practitioner also assigns up to three *domain labels* from a ten-category taxonomy, rates *difficulty* on a five-point scale, and may *flag* the item as incorrect. The difficulty question is subjective: how hard the item was *for that rater*.

The ten-domain taxonomy (Table 1) was defined in consultation with our practitioners, so that it partitions TCM practice the way the field is actually organized. An item can belong to more than one domain: choosing a formula for a gynecological presentation, for example, involves both materia medica and internal medicine. We therefore allow multiple labels, though 82.3% of individual responses use exactly one. The resulting mix follows the licensing curriculum: Internal Medicine is assigned to 41.3% of subset items and Materia Medica to 31.1%, while ENT and Ophthalmology is assigned to only 84 items (1.7%).

This study enrolled 101 licensed practitioners, who produced 15,151 responses over 5,250 distinct items. Of these, 4,801 items received exactly three independent responses. These 4,801 three-rater items form the *comparison*

Table 3: Practitioner-rated difficulty over all 15,151 responses. *Accuracy* is the share of responses at each rating that were correct, and *Time* is the median time spent on the item before submission. Ratings are per response, so an item answered by three practitioners contributes three ratings and may receive a different one from each.

| Rating      | Responses | Share  | Accuracy | Time   |
|-------------|-----------|--------|----------|--------|
| 1 (easiest) | 3,205     | 21.2%  | 74.7%    | 20.3 s |
| 2           | 3,611     | 23.8%  | 69.7%    | 24.5 s |
| 3           | 5,128     | 33.8%  | 59.2%    | 24.9 s |
| 4           | 1,914     | 12.6%  | 49.3%    | 24.0 s |
| 5 (hardest) | 1,293     | 8.5%   | 40.4%    | 23.6 s |
| All         | 15,151    | 100.0% | 62.1%    | 23.5 s |

*subset* on which all human–model analysis is performed; the rest are released but excluded from aggregate human accuracy. We take a *majority vote* over the three annotators, which reduces each item to a single correct/incorrect label and makes practitioner performance directly comparable to a model’s score on the same items. Practitioners spent a median of 23.5 seconds per item, and their difficulty ratings span the full scale (see Table 3). The ratings also track performance: accuracy falls steadily from 74.7% on items rated easiest to 40.4% on items rated hardest.

### 2.3 Evaluation Protocol

We run our testbed on the vLLM inference engine, using greedy decoding at temperature zero with a 16-token budget. From a single generation per item, the harness extracts two predictions. The *generative* prediction  $\hat{y}$  is read from the decoded text: the harness scans it for characters in  $\Sigma$ , then deduplicates them and sorts them into a canonical set. The *log-probability* prediction is taken from the distribution over the first generated token,

$$\hat{y}^{\text{lp}} = \operatorname{argmax}_{c \in \Sigma} \ell(c), \quad m = \ell_{(1)} - \ell_{(2)}, \quad (1)$$

where  $\ell(c)$  is the log-probability of option letter  $c$ , set to  $-\infty$  for letters outside the returned top- $k$  set, and  $m$  is the *margin* between the top two letters.

The second approach separates knowledge from instruction-following. A model that knows the answer but cannot produce the requested format is scored wrong on the generative channel and right on the log-probability one. The gap between the two therefore reveals a formatting failure instead of letting it be mistaken for ignorance. The margin gives us a confidence signal, and it is the basis for the calibration analysis of Appendix B.

Our primary metric is exact match (prediction accuracy),  $\mathbb{1}\{\hat{y} = y\}$ . We apply it the same way to single- and multi-answer items, so a multi-answer item counts as correct only when the model selects every correct option and no incorrect one. For multi-answer items, we also report set-overlap metrics:

$$P = \frac{|\hat{y} \cap y|}{|\hat{y}|}, \quad R = \frac{|\hat{y} \cap y|}{|y|}, \quad F_1 = \frac{2PR}{P+R}. \quad (2)$$

Precision and recall separate two ways of being wrong: a model that names too few options is conservative, and one that names too many is indiscriminate.

## 3 Experiments

We use TCMQA to evaluate 29 instruction-tuned models from 9 families under the protocol of Section 2.3, asking how good current models are at TCM, how they compare against licensed practitioners, and which parts of the field they handle well. Five findings follow.

- The benchmark spans a wide range of difficulty: it separates models across 59 accuracy points, from 18.95% to 77.98%, and no model comes close to saturating it (Section 3.1).
- Pretraining exposure to Chinese-language material matters far more than parameter count: the 12B Gemma is beaten by a Qwen model an eighth its size (Section 3.2).
- Every model loses accuracy on multi-answer items, and exact match hides why: weak models answer as though a single option were wanted (Section 3.3).

Table 4: **Full leaderboard.** Models are grouped by family and ordered by size within each family. *Full* is accuracy on all 38,279 items; *Single* and *Multi* split that by answer type. *Subset* is accuracy on the 4,801 items practitioners also answered, and  $\Delta$  is the difference from their majority vote of 64.94%, in bold where a model wins. All numbers are exact-match percentages.

| Model              | Params | Full  | Single | Multi | Subset | $\Delta$ human |
|--------------------|--------|-------|--------|-------|--------|----------------|
| <b>Gemma2</b>      |        |       |        |       |        |                |
| Gemma-2-2B-It      | 2.61   | 25.09 | 26.44  | 4.78  | 26.74  | −38.20         |
| Gemma-2-9B-It      | 9.24   | 36.93 | 38.72  | 10.10 | 37.78  | −27.16         |
| <b>Gemma3</b>      |        |       |        |       |        |                |
| Gemma-3-270M-It    | 0.27   | 18.95 | 20.20  | 0.00  | 17.93  | −47.01         |
| Gemma-3-1B-It      | 1.00   | 23.82 | 25.17  | 3.39  | 21.85  | −43.09         |
| Gemma-3-4B-It      | 4.30   | 30.19 | 31.54  | 9.97  | 30.31  | −34.63         |
| Gemma-3-12B-It     | 12.19  | 39.60 | 41.03  | 18.11 | 40.93  | −24.01         |
| <b>Gemma4</b>      |        |       |        |       |        |                |
| Gemma-4-E2B-It     | 5.12   | 33.43 | 34.68  | 14.54 | 36.80  | −28.14         |
| Gemma-4-E4B-It     | 8.00   | 44.69 | 46.65  | 15.09 | 46.43  | −18.51         |
| <b>OLMo2</b>       |        |       |        |       |        |                |
| OLMo-2-1B-Inst.    | 1.48   | 21.05 | 22.45  | 0.00  | 20.83  | −44.11         |
| OLMo-2-7B-Inst.    | 7.30   | 24.36 | 25.72  | 3.86  | 25.83  | −39.11         |
| OLMo-2-13B-Inst.   | 13.72  | 26.33 | 27.27  | 12.11 | 26.37  | −38.57         |
| <b>OLMo3</b>       |        |       |        |       |        |                |
| OLMo-3-7B-Inst.    | 7.30   | 26.82 | 28.18  | 6.33  | 26.85  | −38.09         |
| <b>Qwen1.5</b>     |        |       |        |       |        |                |
| Qwen1.5-0.5B-Chat  | 0.62   | 23.99 | 25.33  | 3.90  | 23.89  | −41.05         |
| Qwen1.5-1.8B-Chat  | 1.84   | 38.53 | 40.63  | 6.96  | 44.37  | −20.57         |
| Qwen1.5-4B-Chat    | 3.95   | 49.85 | 52.31  | 12.82 | 56.61  | −8.33          |
| Qwen1.5-7B-Chat    | 7.72   | 61.23 | 63.78  | 22.88 | 70.71  | +5.77          |
| Qwen1.5-14B-Chat   | 14.17  | 67.89 | 69.97  | 36.50 | 78.11  | +13.17         |
| <b>Qwen2</b>       |        |       |        |       |        |                |
| Qwen2-0.5B-Inst.   | 0.49   | 35.15 | 36.49  | 14.96 | 39.49  | −25.45         |
| Qwen2-1.5B-Inst.   | 1.54   | 60.81 | 63.70  | 17.23 | 71.88  | +6.94          |
| Qwen2-7B-Inst.     | 7.62   | 75.02 | 76.66  | 50.29 | 86.73  | +21.79         |
| <b>Qwen2.5</b>     |        |       |        |       |        |                |
| Qwen2.5-0.5B-Inst. | 0.49   | 29.77 | 31.43  | 4.74  | 29.70  | −35.24         |
| Qwen2.5-1.5B-Inst. | 1.54   | 52.31 | 54.28  | 22.55 | 57.78  | −7.16          |
| Qwen2.5-3B-Inst.   | 3.09   | 59.27 | 61.78  | 21.54 | 67.32  | +2.38          |
| Qwen2.5-7B-Inst.   | 7.62   | 71.59 | 73.58  | 41.66 | 81.48  | +16.54         |
| Qwen2.5-14B-Inst.  | 14.77  | 77.98 | 79.51  | 54.99 | 86.38  | +21.44         |
| <b>Qwen3.5</b>     |        |       |        |       |        |                |
| Qwen3.5-0.8B       | 0.87   | 36.82 | 38.27  | 14.92 | 41.12  | −23.82         |
| Qwen3.5-2B         | 2.27   | 50.77 | 53.82  | 4.90  | 56.51  | −8.43          |
| Qwen3.5-4B         | 4.66   | 67.37 | 69.59  | 34.03 | 73.57  | +8.63          |
| Qwen3.5-9B         | 9.65   | 75.02 | 77.06  | 44.30 | 81.44  | +16.50         |

- Nine of the 29 models outperform the practitioner majority vote of 64.94% on the shared items, and all nine come from the Qwen family (Section 3.4).
- Practitioners and models do not find the same items hard: model accuracy is flat across practitioner consensus and practitioner-rated difficulty (Section 3.5).

### 3.1 Overall Performance

**Setup.** We first ask whether our benchmark separates models at all. We evaluate 29 instruction-tuned models from 9 families, ranging from 0.27B to 14.77B parameters, under the protocol of Section 2.3. Most are small or mid-sized open-weight models, the kind that ordinary users can run locally, which matters in the medical domain, where privacy concerns push users toward local models. We include both Chinese- and Western-developed models.

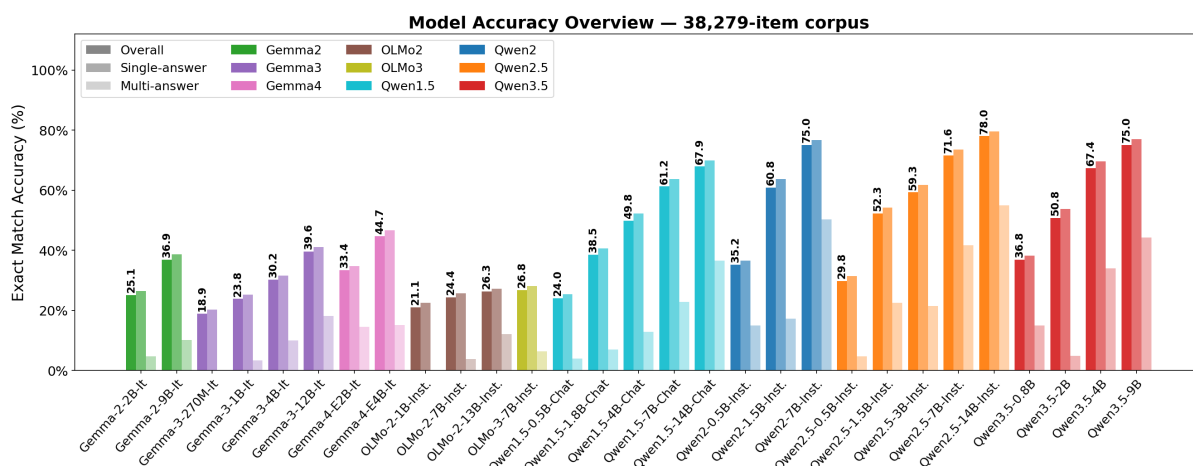


Figure 1: **Model accuracy by answer type.** Overall, single-answer, and multi-answer accuracy for all 29 models, grouped by family. The multi-answer bar is shorter than the single-answer bar for all 29 models, and the gap grows as models get weaker, so the overall score hides this weakness.

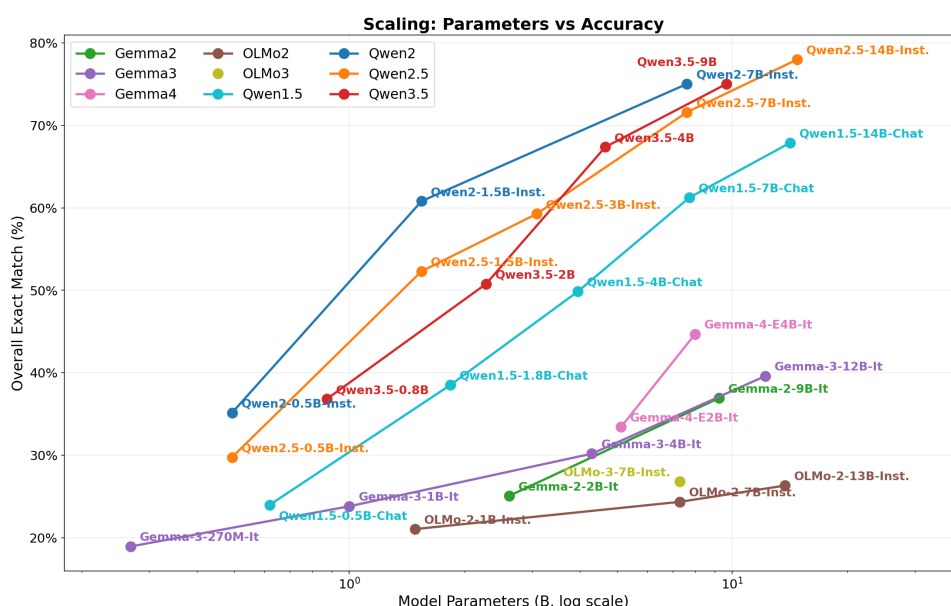


Figure 2: **Accuracy against model size.** Accuracy against parameter count on a log axis, one line per family. The Qwen lines sit in a band of their own, and the gap to Gemma and OLMo grows with size instead of closing. Note also that Qwen2 beats Qwen2.5 at 0.5B, 1.5B, and 7B; Qwen2.5 wins only at 14B, where Qwen2 has no model, so Qwen2.5's top spot comes from being bigger rather than better.

All 29 are scored by exact match on the full corpus of 38,279 items, reported overall and split by answer type.

**Results.** Table 4 and Figure 1 report accuracy for all 29 models. Two findings stand out. First, every model does worse on multi-answer questions than on single-answer ones, and the gap widens as models get weaker. Second, the benchmark spreads models out: scores range from 18.95% to 77.98%. The best model, Qwen2.5-14B-Inst., gets about four questions in five right, while the worst, Gemma-3-270M-It, performs no better than random guessing. Two further patterns preview the sections that follow: the Qwen models (Qwen3.5, Qwen2.5, Qwen2, and even Qwen1.5) sit at the frontier of the leaderboard, ahead of Western-developed families such as Gemma and OLMo, and nine of the 29 models outperform the practitioner reference—all nine of them Qwen.

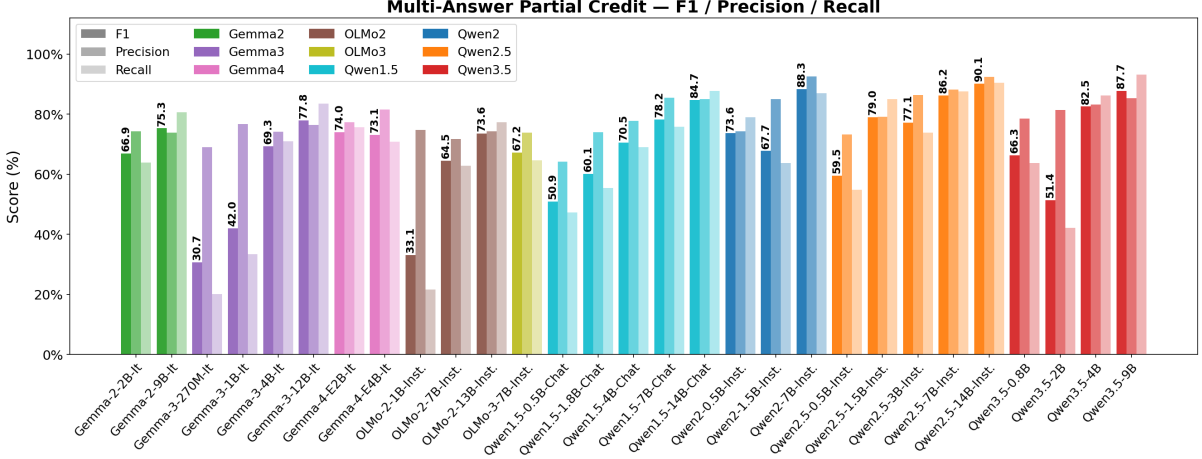


Figure 3: **Precision, recall and  $F_1$  on multi-answer items.** Set-overlap metrics on the 2,386 multi-answer items. Precision beats recall for most models. That means they name too few options and treat these questions as single-answer.

### 3.2 Training Data Matters More Than Size

**Setup.** Section 3.1 established that the benchmark separates models across 59 accuracy points. We now ask what explains that spread. We plot accuracy against parameter count for all 29 models on a logarithmic axis, one line per family (Figure 2). The families span 0.27B to 14.77B parameters, and six of the nine cover at least three sizes, so both the within-family and the across-family comparison are available over a usable range.

**Results.** First, within every family, accuracy rises monotonically with size: once the training data is held fixed, scale buys real gains. Across families, however, scale stops predicting accuracy. Gemma-3-12B-It reaches 39.60%, while Qwen2-1.5B-Inst., eight times smaller, reaches 60.81%; with the single exception of Qwen1.5-1.8B-Chat, every Qwen model at 1.5B or above beats the 12B Gemma. The OLMo family never passes 27% at any size we tried, and growing from 1B to 13B gains it only about five points. Every model that outperforms the practitioner baseline is a Qwen model. Second, newer generations narrow the gap without closing it. Qwen3.5 climbs steadily from 36.82% to 75.02% across our four sizes, and Qwen3.5-4B matches Qwen1.5-14B-Chat with a third of the parameters. Gemma-4-E4B-It is the strongest Gemma we measured, and it still loses to Qwen3.5-2B, a model  $3.5\times$  smaller from a family two generations older. Third, progress is not uniformly forward, either: Qwen2.5 scores below Qwen2 at all three sizes where both exist. The most plausible explanation is the composition of the pretraining corpus, since the families that do well are the ones trained on far more Chinese text.

### 3.3 Answer Sets in Multi-Answer Items

**Setup.** Section 3.1 showed that every model loses accuracy on multi-answer items. We look more closely at those items here. We score the 2,386 multi-answer items with precision, recall, and  $F_1$  alongside exact match, and record  $|\hat{y}|$ , the mean number of options a model names, against a mean answer size of 3.50.

**Results.** Multi-answer items hurt far more than their 6.2% share suggests (Table 5). The best model falls from 79.51% on single-answer items to 54.99% on multi-answer ones, and models below 55% overall often land in single digits. The failures are systematic, and they run in two directions (Figure 3). Strong models answer about the right number of options: Qwen2.5-14B-Inst. answers 3.40 with precision and recall both above 90%, so it picks a slightly wrong set rather than a set of the wrong size. Weak models name too few options or too many. Qwen3.5-2B names 1.84, buying a precision of 81.3% at a recall of 42.2%; Gemma-3-270M-It names 0.90, as though a single option were wanted; and Gemma-3-12B-It overshoots at 3.82. Exact match scores these habits identically, so it understates what a model that names too few options actually knows.

### 3.4 Comparing Against Licensed Practitioners

**Setup.** We now turn to the human reference, and ask how open-weight models compare against the people licensed to practice. We restrict to the 4,801 items that three practitioners each answered and reduce the three

Table 5: **Multi-answer decomposition.** Results on the 2,386 multi-answer items.  $|\hat{y}|$  is the mean number of options predicted; the answers contain 3.50 on average. Precision above recall means the model named too few options, recall above precision means too many.

| Model             | Single EM | Multi EM | P    | R    | F <sub>1</sub> | $ \hat{y} $ |
|-------------------|-----------|----------|------|------|----------------|-------------|
| Qwen2.5-14B-Inst. | 79.5      | 55.0     | 92.4 | 90.4 | 90.1           | 3.40        |
| Qwen2-7B-Inst.    | 76.7      | 50.3     | 92.5 | 86.9 | 88.3           | 3.25        |
| Qwen3.5-9B        | 77.1      | 44.3     | 85.3 | 93.2 | 87.7           | 3.84        |
| Qwen3.5-2B        | 53.8      | 4.9      | 81.3 | 42.2 | 51.4           | 1.84        |
| Gemma-3-12B-It    | 41.0      | 18.1     | 76.3 | 83.4 | 77.8           | 3.82        |
| OLMo-2-13B-Inst.  | 27.3      | 12.1     | 74.2 | 77.2 | 73.6           | 3.61        |
| Gemma-3-270M-It   | 20.2      | 0.0      | 69.0 | 20.2 | 30.7           | 0.90        |

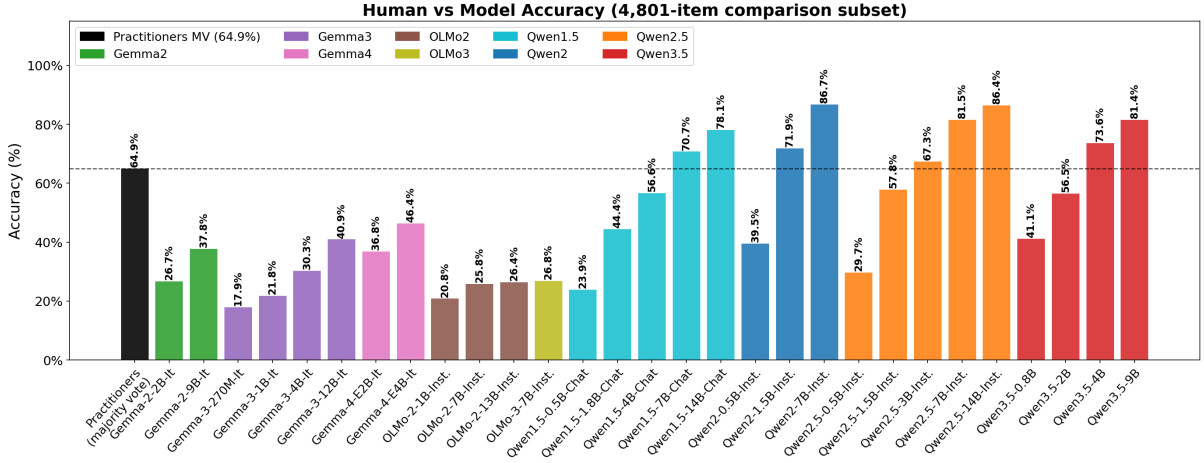


Figure 4: **Models against the practitioner reference.** Accuracy on the 4,801 shared items: the practitioner majority vote against all 29 models, with the practitioner score as a dashed line. Nine models clear the line and all nine are Qwen. Every Gemma and OLMo falls short, including models larger than the Qwen ones that succeed.

responses to a majority vote, which yields a single correct-or-incorrect label per item at 64.94% accuracy. All 29 models are scored on those same items, so both sides see identical questions, which is what makes the comparison fair even though the subset is smaller than the full corpus.

**Results.** Nine of 29 models beat the practitioner majority vote (Figure 4). Qwen2-7B-Inst. leads at 86.73%, ahead by 21.79 points, and the margin stays positive down to Qwen2.5-3B-Inst. All nine are Qwen models, and the twenty below the line include every Gemma and every OLMo we tested, several of them larger than the Qwen models above the line.

The advantage does not come from one weak area (Figure 5). The three best models beat the practitioners in all ten domains. What changes from domain to domain is the size of the gap, and it changes because practitioner scores do. Practitioners are weakest on Regulation at 50.9%, where the best model reaches 87.9%, and strongest on Diagnosis and Patterns at 73.0% and Foundational Theory at 71.3%, where the same model is only about 15 points ahead. Model accuracy hardly moves across domains, so the models are furthest ahead on material that is mostly memorization and closest on material that needs clinical reasoning.

### 3.5 Models and Practitioners Find Different Things Hard

**Setup.** Nine models outscore the practitioners, but a higher score need not mean the same ability. We therefore ask whether the two sides (models and humans) find the same items hard. On the shared subset we relate model accuracy to two practitioner signals, how many of the three answered each item correctly and the difficulty they rated it before learning the outcome, and we measure item-level agreement with Cohen’s  $\kappa$ . Flat accuracy against the first two signals would mean the two kinds of difficulty are unrelated, but only if the practitioner signal carries information at all. We test that first against Binomial(3, 0.6248), the distribution of correct counts had the three practitioners answered independently at their observed rate.

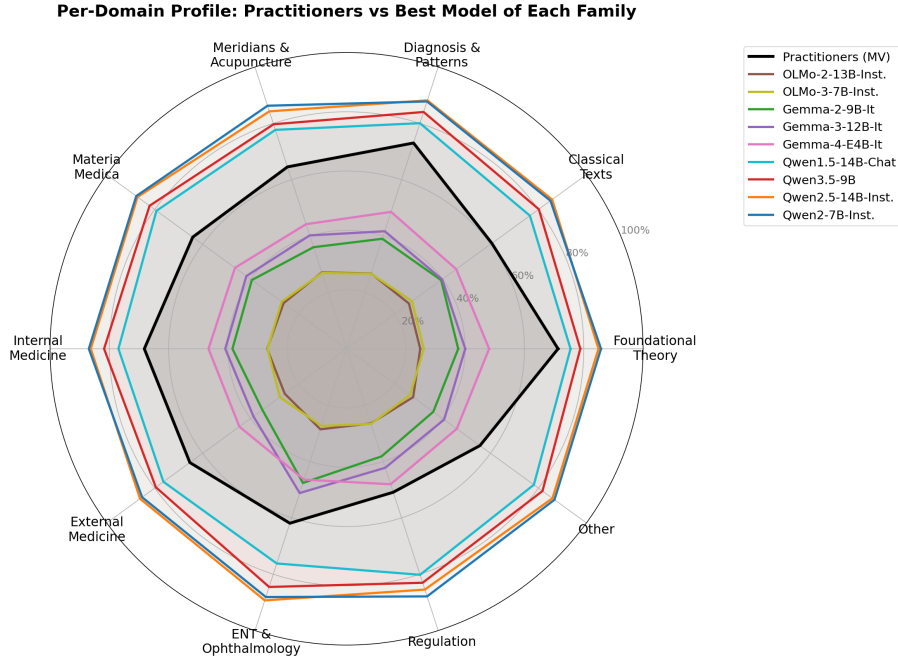


Figure 5: **Per-domain accuracy.** Accuracy across the ten practitioner-assigned domains: the practitioner majority vote against the best model of each family. The order is the same on every axis. The three Qwen contours sit outside the practitioner contour everywhere, and every Gemma and OLMo contour sits inside it. The model contours are nearly circular while the practitioner contour is lopsided, so the gap is shaped by the human side.

Table 6: **Practitioner agreement against independence.** Agreement is far stronger than independent answering would produce. Expected counts come from Binomial(3, 0.6248) over the 4,801 items with three raters.

| Practitioners correct                              | Observed | If independent | Ratio |
|----------------------------------------------------|----------|----------------|-------|
| 0 of 3                                             | 616      | 253.6          | 2.43× |
| 1 of 3                                             | 1,067    | 1,266.8        | 0.84× |
| 2 of 3                                             | 1,422    | 2,109.6        | 0.67× |
| 3 of 3                                             | 1,696    | 1,171.0        | 1.45× |
| $\chi^2 = 1009.0$ (3 df), Fleiss' $\kappa = 0.263$ |          |                |       |

**Results.** The practitioner signal is real (Table 6). Items that all three practitioners miss are 2.43 times more common than independence predicts and items all three answer are 1.45 times more common, giving  $\chi^2 = 1009.0$  on three degrees of freedom against a critical value of 16.27. Practitioners agree about which questions are hard, far more than chance would allow.

Model accuracy ought to follow that signal, and it does not (Figure 6). On the 616 items every practitioner missed, Qwen2-7B-Inst. scores 87.34%, slightly *above* its 86.73% subset average; on the items every practitioner answered it scores 86.50%. The same flatness holds for all 29 models, down to those near 20%. The difficulty ratings agree (Figure 7): practitioner accuracy falls 57 points, from 89.9% at level 1 to 32.6% at level 5, while over the same range Qwen2-7B-Inst. loses ten points, Qwen3.5-9B under two, and Gemma-3-12B-It stays between 39.2% and 42.7% throughout. Gemma-4-E4B-It is the one real exception, and its 17.5-point decline is still a third of the practitioners'. Models overtake practitioners on hard items because the practitioners fall away, not because the models hold up.

Item-level agreement points the same way. Cohen's  $\kappa$  between a model and the practitioner majority runs from  $-0.017$  to  $+0.023$  across all 29 models, indistinguishable from zero and with no trend in model quality, while between pairs of models it runs from 0.06 to 0.23, up to ten times higher. Models err like each other rather than like the practitioners, which points to habits picked up in pretraining rather than to intrinsically hard items. Figure 8 shows what that costs. For Qwen2-7B-Inst., both sides answer 56.5% of the items and neither answers 4.8%, but the model alone answers 30.2% and the practitioners alone 8.4%. If a model were simply a better practitioner, that last cell would shrink toward zero as accuracy rose. It does not: across the leaderboard, practitioner-only wins fall

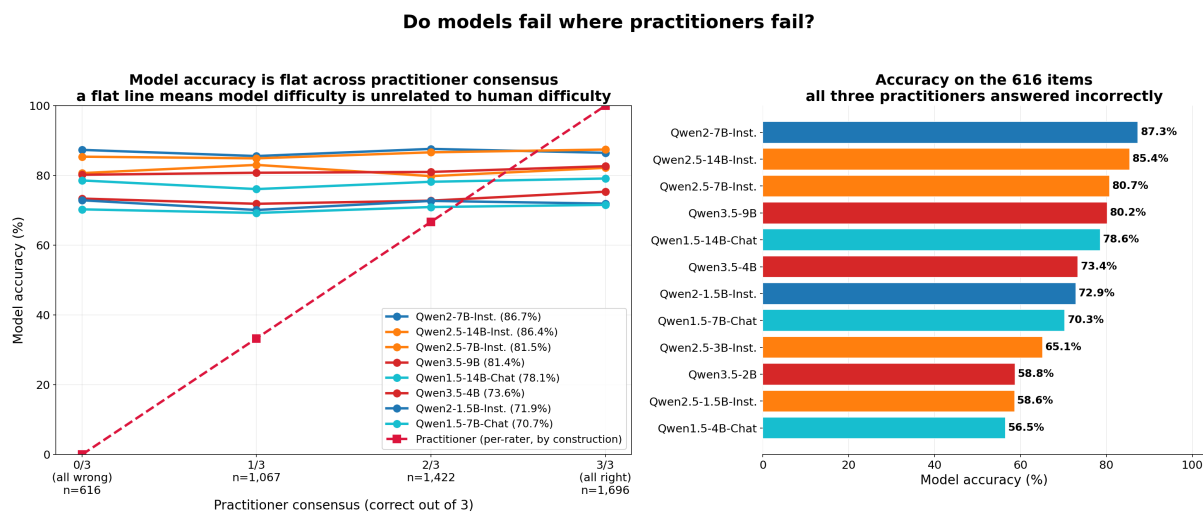


Figure 6: **Model accuracy by practitioner consensus.** Left: model accuracy by how many of the three practitioners answered correctly. Right: accuracy on the 616 items all three got wrong. The model lines are flat, and the best model does slightly better on the questions every practitioner missed; the dashed practitioner line runs from 0% to 100% by construction. Because practitioner agreement is real rather than noise (Table 6), flat lines mean the two difficulties are unrelated.

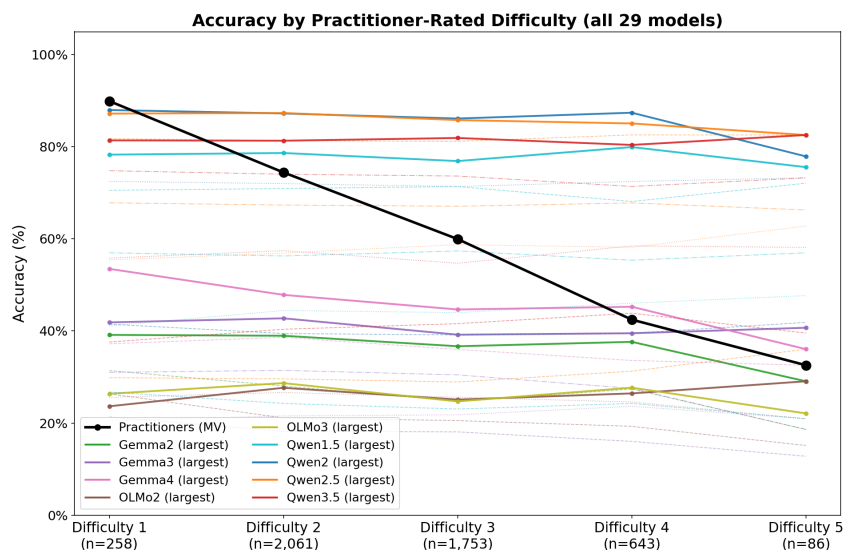


Figure 7: **Accuracy across the difficulty scale.** All 29 models plus the practitioner line, over the practitioners' five-point scale. Practitioner accuracy falls 57 points while almost every model line stays flat: at level 1 the practitioners beat all but two models, and by level 5 they are below Gemma-4-E4B-It. The ratings were given before anyone knew the outcome (Section 2.2), so this is separate evidence rather than a restatement of Figure 6. Level 5 holds only 86 items, so read the last point loosely.

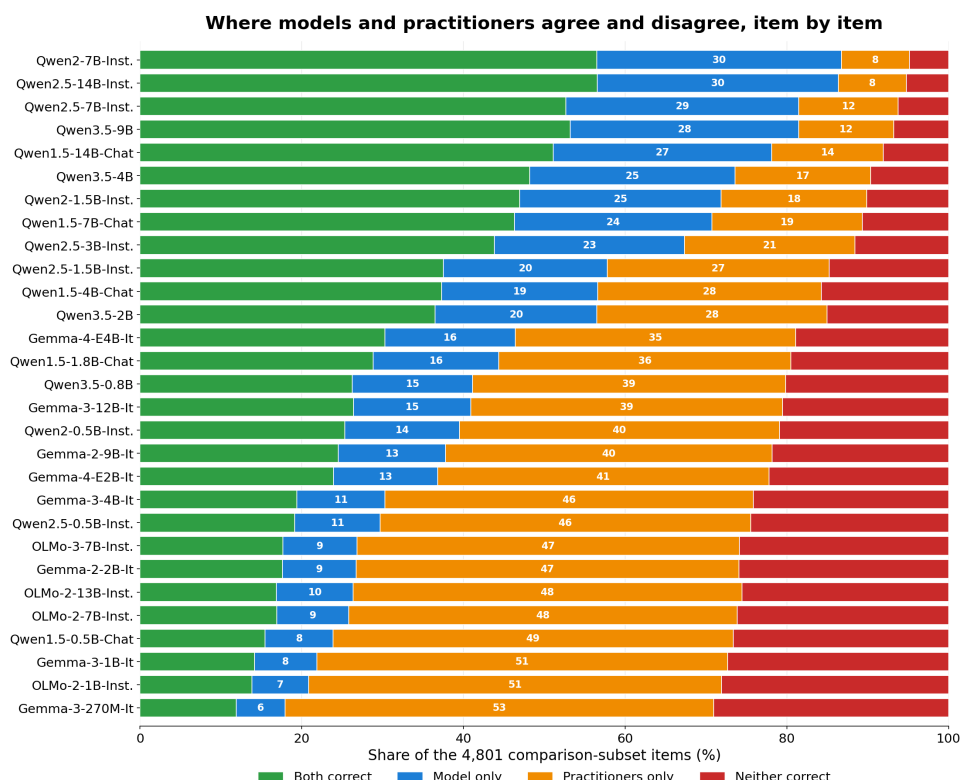


Figure 8: **Item-by-item outcomes.** Outcomes against the practitioner majority vote for all 29 models, ordered by accuracy; the numbers inside the bars are the two disagreement cells as a share of the 4,801 items. Going from the worst model to the best raises model-only wins from 6% to 30%, but practitioner-only wins fall from 53% to 8% and stop there. The orange band never closes, so a single ranked score hides more than it shows.

from 53% to 8% and then stop, so even our best model leaves several hundred items that the practitioners answer and it misses.

## 4 Conclusion

TCMQA pairs 38,279 questions from Chinese TCM licensing examinations with 15,151 responses from 101 licensed practitioners. Across 29 instruction-tuned models from 9 families, the benchmark separates the field by 59 accuracy points and no model comes close to saturating it. What a model was pretrained on predicts TCM ability far better than how large it is. Nine models clear the practitioner majority vote of 64.94%, the best by 21.79 points, and every one of them comes from a single Chinese-pretrained family; Western-pretrained models several times larger fall more than twenty points short.

Practitioners agree about which items are hard, yet model accuracy is flat across their consensus and nearly flat across their difficulty ratings. Agreement with the practitioner majority is indistinguishable from zero for all 29 models, while models agree with each other up to ten times more often, and even the leading model misses several hundred items the practitioners answer. The models are furthest ahead on Regulation, where the material is largely memorization, and closest on Diagnosis and Patterns.

## References

- [1] Bean, A. M.; Payne, R. E.; Parsons, G.; Kirk, H. R.; Ciro, J.; Mosquera-Gómez, R.; Hincapié M, S.; Ekanayaka, A. S.; Tarassenko, L.; Rocher, L.; others Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine* **2026**, 32, 609–615.
- [2] Meng, X.; Yan, X.; Zhang, K.; Liu, D.; Cui, X.; Yang, Y.; Zhang, M.; Cao, C.; Wang, J.; Wang, X.; others The application of large language models in medicine: A scoping review. *Iscience* **2024**, 27.
- [3] Kunze, K. N.; Nwachukwu, B. U.; Cote, M. P.; Ramkumar, P. N. Large language models applied to health care tasks may improve clinical efficiency, value of care rendered, research, and medical education. *Arthroscopy* **2025**, 41, 547–556.
- [4] Mingole, B.; Majumdar, A.; Choudhury, F. A.; Kraschnewski, J. L.; Sundar, S. S.; Yadav, A. Dr. GPT Will See You Now, but Should It? Exploring the Benefits and Harms of Large Language Models in Medical Diagnosis using Crowdsourced Clinical Cases. *arXiv preprint arXiv:2506.13805* **2025**,
- [5] Jindal, J.; Bedi, S.; Swaminathan, A.; Wornow, M.; Fries, J.; Chaurasia, A.; Shah, N. Large language models in healthcare: Are we there yet? *Stanford HAI* **2024**,
- [6] Roustan, D.; Bastardot, F. The clinicians’ guide to large language models: a general perspective with a focus on hallucinations. *Interactive journal of medical research* **2025**, 14, e59823.
- [7] Busch, F.; Hoffmann, L.; Rueger, C.; Van Dijk, E. H.; Kader, R.; Ortiz-Prado, E.; Makowski, M. R.; Saba, L.; Hadamitzky, M.; Kather, J. N.; others Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine* **2025**, 5, 26.
- [8] Wu, E.; Wu, K.; Hom, J.; Yi, P. H.; Zhang, A.; Lozano, A.; Nirschl, J.; Tangney, J.; Byram, K.; Dymm, B.; others MedArena: Comparing LLMs for Medicine-in-the-Wild Clinician Preferences. *arXiv preprint arXiv:2603.15677* **2026**,
- [9] Pal, A.; Umapathi, L. K.; Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. Conference on health, inference, and learning. 2022; pp 248–260.
- [10] Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **2021**, 11, 6421.
- [11] Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019; pp 2567–2577.
- [12] Liu, Y.; Yuan, Y.; Yan, K.; Li, Y.; Sacca, V.; Hodges, S.; Cannistra, M.; Jeong, P.; Wu, J.; Kong, J. Evaluating the role of large language models in traditional Chinese medicine diagnosis and treatment recommendations. *NPJ digital medicine* **2025**, 8, 466.
- [13] Han, C.; Yang, G.; Li, H.; Zhu, L.; Feng, M. Tuning and clinical application of large language models in Traditional Chinese Medicine: scoping review. *Chinese Medicine* **2026**, 21, 71.
- [14] Li, Y.-X.; Elnaffar, S.; Chen, H.-Y.; Chen, N.-J.; Lai, P.-Y.; Li, N.-Q.; Chong, Y.; Qiao, J.; Liu, T.; Peng, Z.-B.; others An LLM method for understanding traditional Chinese medicine: Mechanism exploration and innovative application. *IEEE Journal of Biomedical and Health Informatics* **2025**,
- [15] Yue, W.; Wang, X.; Zhu, W.; Guan, M.; Zheng, H.; Wang, P.; Sun, C.; Ma, X. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine. *arXiv preprint arXiv:2406.01126* **2024**,
- [16] Cheng, Z.; Lu, Y.; Ye, H.; Liu, Z.; Wang, M.; Liu, J.; Li, Z.; Fan, W.; Guo, Y.; Fu, R.; others Tcm-eval: An expert-level dynamic and extensible benchmark for traditional chinese medicine. *arXiv preprint arXiv:2511.07148* **2025**,
- [17] Gemma Team; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; others Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786* **2025**,
- [18] Qwen Team Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* **2024**,
- [19] Olmo, T.; Ettinger, A.; Bertsch, A.; Kuehl, B.; Graham, D.; Heineman, D.; Groeneveld, D.; Brahman, F.; Timbers, F.; Ivison, H.; others Olmo 3. *arXiv preprint arXiv:2512.13961* **2025**,

[20] OLMo Team; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; others 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656* **2025**,

## A Annotation Interface

The human reference of Section 2.2 was collected through a web application built for this study. A practitioner signs in, is assigned a private block of 150 items, and sees one item per screen. The screen carries four controls, shown one at a time in Figures 9–12, and a single submission records all of them together before advancing to the next unanswered item.

問題 23/150 (单选题).

腸風脏毒所致的便前或便后出血，可選用

- ☐ 黄土湯
- ☐ 槐花散
- ☐ 白頭翁湯
- ☐ 四物湯
- ☐ 歸脾湯

請選擇正確答案

此問題的類別你覺得是？(至多三種)

|                                   |                                               |                                            |                                            |
|-----------------------------------|-----------------------------------------------|--------------------------------------------|--------------------------------------------|
| <input type="checkbox"/> 【中醫基礎理論】 | <input type="checkbox"/> 【中醫經典文獻】             | <input type="checkbox"/> 【診斷與辨證學】          | <input type="checkbox"/> 【經絡與針灸學】          |
| <input type="checkbox"/> 【中藥本草學】  | <input type="checkbox"/> 【中醫內科學】 (含婦科兒科、精神志病) | <input type="checkbox"/> 【中醫外科學】 (含皮膚科、瘡科) | <input type="checkbox"/> 【中醫五官科學】 (含眼耳鼻咽喉) |
| <input type="checkbox"/> 【中醫法規】   | <input type="checkbox"/> 【其他】                 |                                            |                                            |

對你來說問題困難度？(1~5)

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

查看問題快速鍵和類別描述

下一個

Figure 9: **Answering an item.** The item is presented in its original Chinese with its option list, one item per screen, and the header shows position within the practitioner’s 150-item block and whether the item is single- or multi-answer. Models see the same item text under the protocol of Section 2.3, so both sides are answering identical questions.

問題 23/150 (单选题).

腸風脏毒所致的便前或便后出血，可選用

- ☐ 黄土湯
- ☐ 槐花散
- ☐ 白頭翁湯
- ☐ 四物湯
- ☐ 歸脾湯

請選擇所屬領域 (至多三項)

此問題的類別你覺得是？(至多三種)

|                                   |                                               |                                            |                                            |
|-----------------------------------|-----------------------------------------------|--------------------------------------------|--------------------------------------------|
| <input type="checkbox"/> 【中醫基礎理論】 | <input type="checkbox"/> 【中醫經典文獻】             | <input type="checkbox"/> 【診斷與辨證學】          | <input type="checkbox"/> 【經絡與針灸學】          |
| <input type="checkbox"/> 【中藥本草學】  | <input type="checkbox"/> 【中醫內科學】 (含婦科兒科、精神志病) | <input type="checkbox"/> 【中醫外科學】 (含皮膚科、瘡科) | <input type="checkbox"/> 【中醫五官科學】 (含眼耳鼻咽喉) |
| <input type="checkbox"/> 【中醫法規】   | <input type="checkbox"/> 【其他】                 |                                            |                                            |

對你來說問題困難度？(1~5)

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

查看問題快速鍵和類別描述

下一個

Figure 10: **Assigning domain labels.** The practitioner tags the item with up to three of the ten domains of Table 1, plus 其他 (Other) for items the taxonomy does not cover. Multiple labels are allowed because one item can span domains, though 82.3% of responses use exactly one. These labels produce the per-domain comparison of Figure 5.

問題 23/150 (单选题).

腸風飩毒所致的便前或便后出血，可選用

☐ 黃土湯  
☐ 槐花散  
☐ 白頭翁湯  
☐ 四物湯  
☐ 歸脾湯

此問題的類別你覺得是？(至多三種)

☐ 【中醫基礎理論】  
☐ 【中醫經典文獻】  
☐ 【中醫內科學】(含婦科兒科神志病)  
☐ 【中醫法理】  
☐ 【診斷與辨證學】  
☐ 【中醫外科學】(含皮膚科、瘡科)  
☐ 【經絡與針灸學】  
☐ 【中醫五官科學】(含眼耳鼻咽喉)

對你來說問題困難度？(1-5)

☐ 1  
☐ 2  
☐ 3  
☐ 4  
☐ 5

本題難易度 (5 是最難)

查看錯誤快捷鍵和類別描述

下一個

Figure 11: **Rating difficulty.** A five-point scale, 5 being hardest. The question asks how hard the item was *for that rater* rather than how hard it is in general, so the ratings are subjective by design. They are collected on the same screen as the answer and before any outcome is revealed, which is what lets Figure 7 treat them as independent evidence.

問題 23/150 (单选题).

腸風飩毒所致的便前或便后出血，可選用

☐ 黃土湯  
☐ 槐花散  
☐ 白頭翁湯  
☐ 四物湯  
☐ 歸脾湯

此問題的類別你覺得是？(至多三種)

☐ 【中醫基礎理論】  
☐ 【中醫經典文獻】  
☐ 【中醫內科學】(含婦科兒科神志病)  
☐ 【中醫法理】  
☐ 【診斷與辨證學】  
☐ 【中醫外科學】(含皮膚科、瘡科)  
☐ 【經絡與針灸學】  
☐ 【中醫五官科學】(含眼耳鼻咽喉)

對你來說問題困難度？(1-5)

☐ 1  
☐ 2  
☐ 3  
☐ 4  
☐ 5

查看錯誤快捷鍵和類別描述

下一個

問題報錯 暫停一下

若題目有問題，可以在這裡報錯

Figure 12: **Flagging and pausing.** 問題報錯 flags an item the practitioner believes is defective, and 暫停一下 pauses the block. The flag gives a practitioner somewhere to report a bad item instead of guessing at it, and the pause keeps an interruption from being recorded as a slow or careless response; timing is measured per item, and the median is 23.5 seconds.

## B Model Confidence and Calibration

Section 3 ranks models by how often they are right. Here we ask a second question: when a model is wrong, does it know?

**Setup.** We take the margin  $m$  of Equation (1) as the confidence signal. For each model we group its items by margin and measure accuracy within each group, which gives one calibration curve per model (Figure 13). We then reduce each model to a single number, the gap between its mean confidence on correct answers and on incorrect ones, with confidence capped at the 99th percentile to keep a few extreme margins from dominating (Figure 14). A large gap means the model’s confidence is worth listening to; a gap near zero means it is not.

**Results.** Two findings stand out. First, confidence points the right way in every family: accuracy rises with the margin every time. What changes is the scale. Qwen2.5 and Gemma4 spread out over margins as large as 20 nats and get nearly everything right at the top end, while Qwen3.5 squeezes its whole curve into a much narrower range, so a margin of 3 means different things for different models. Second, calibration does not follow from accuracy. Qwen2.5-14B-Inst. reaches a confidence gap of 9.5, while Qwen3.5-9B manages only 3.1 at similar accuracy, and Gemma-3-1B-It and OLMo-2-13B-Inst. fall below 1.0. Gemma-3-1B-It and Gemma-3-4B-It are the worst case:

inaccurate and confident all the same. A model with a small gap should not be trusted to decide when to answer, however good its overall score is.

### Calibration: Confidence Margin vs Accuracy

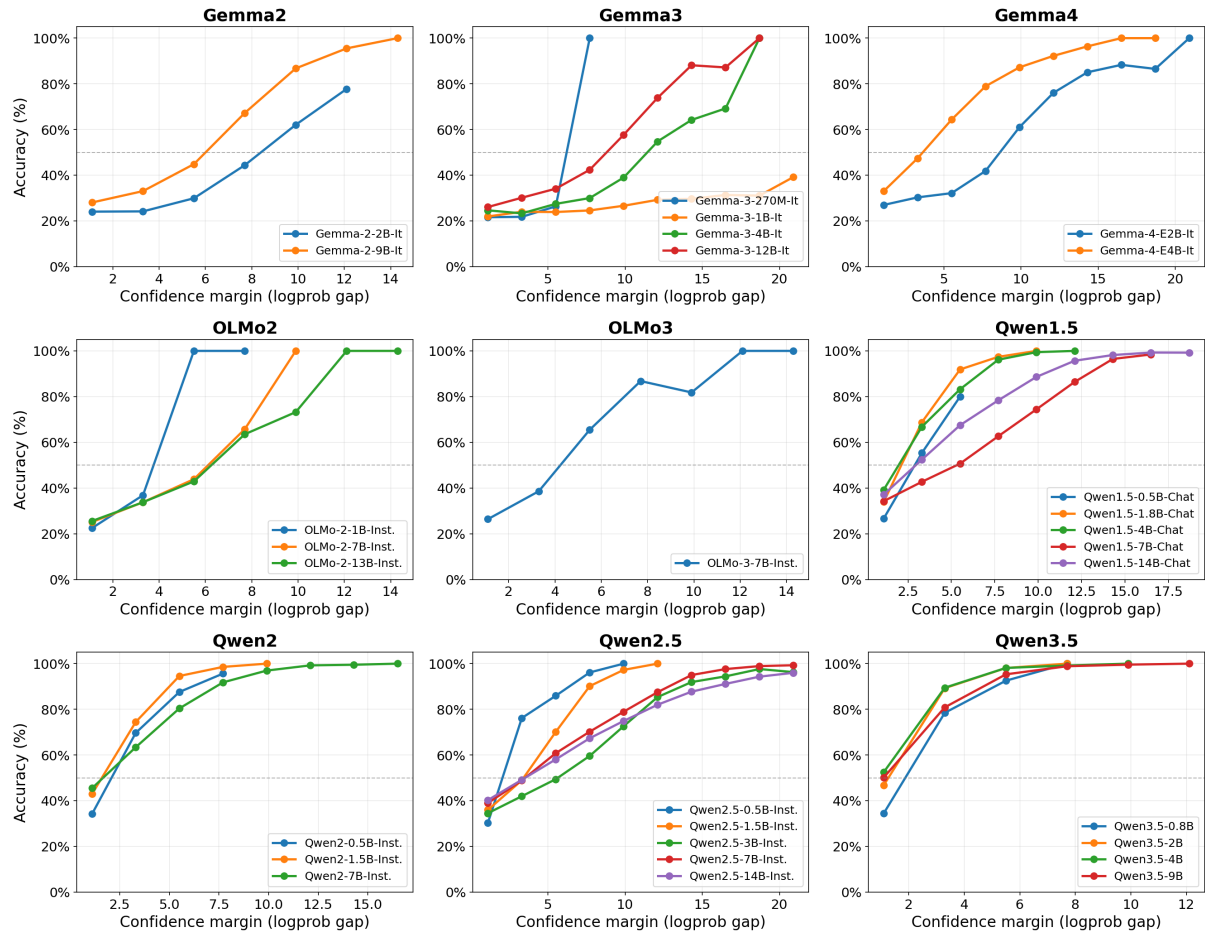


Figure 13: **Accuracy by confidence margin.** Accuracy grouped by log-probability margin, one panel per family. More confidence really does mean more accuracy in every family, but the useful range varies by close to a factor of ten, so margins from different models are not comparable. The panels have independent  $x$ -axes, so compare the ranges rather than the slopes. Qwen3.5 is the narrowest, fitting its whole curve into a range where the other families are still climbing.

### Overconfidence Analysis (single-answer items)

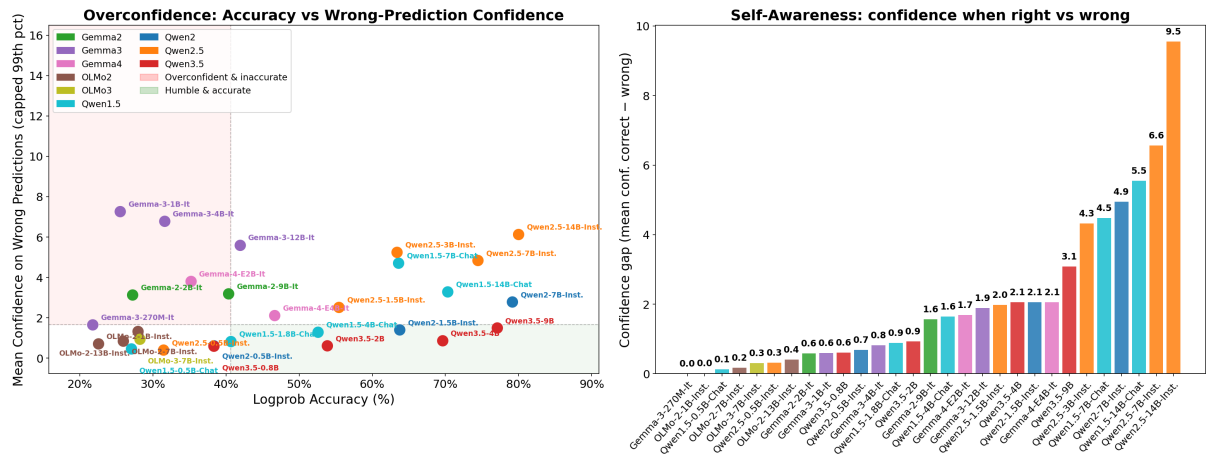


Figure 14: **Overconfidence.** Left: accuracy against average confidence on *wrong* answers, split at the medians. Right: the confidence gap between right and wrong answers, with confidence capped at the 99th percentile. The top-left corner is where models are inaccurate and sure of themselves, and two Gemma3 models sit there. The right panel shows this does not follow from accuracy: Qwen2.5-14B-Inst. and Qwen3.5-9B score similarly but separate right from wrong answers by 9.5 and 3.1.