

# Med-GRADE: Medical Grading and Rubric-based Assessment of Doctor-Patient Encounters

Yidong Jiang<sup>1\*</sup> Chao Ding<sup>2,3\*</sup> Renjie Lu<sup>3,4</sup> Rongzhao Zhang<sup>2</sup> Yankai Jiang<sup>2</sup>  
Tianbin Li<sup>2</sup> Junzhi Ning<sup>2</sup> Chengze Li<sup>5</sup> Haoyu Zhang<sup>6</sup>  
Jinru Ding<sup>2†</sup> Pengcheng Chen<sup>7‡</sup>

<sup>1</sup>Tongji University <sup>2</sup>Shanghai Artificial Intelligence Laboratory

<sup>3</sup>Shanghai University of Traditional Chinese Medicine and Pharmacology <sup>4</sup>Longhua Hospital

<sup>5</sup>University of Illinois at Chicago <sup>6</sup>Shanghai Innovation Institute <sup>7</sup>ByteDance

\*Equal Contribution †Project leader ‡Corresponding author

## Abstract

Large language model (LLM)-based judges are increasingly used to scale the evaluation and optimization of open-ended medical dialogue systems, making their reliability critical in this high-stakes setting. However, existing medical benchmarks mainly assess their medical knowledge, while general judge benchmarks lack clinically grounded consultation criteria. To address this gap, we introduce Med-GRADE, a benchmark for evaluating LLMs as medical dialogue judges. It contains 2,076 real-world patient-LLM dialogues annotated by 56 physicians with a 23-item rubric adapted from Mini-CEX, covering Medical Interviewing (MI), Humanistic Care (HC), and Diagnosis and Treatment Management (DTM). Experiments on 17 LLMs show that current LLM judges achieve only moderate agreement, with the best model reaching 64.89% Macro-F1 and 68.28% Hamming Accuracy. Besides, more complex prompts may even harm performance: CoT disrupts parsing, while few-shot and self-refine introduce item-level instability. Data and code are available at <https://github.com/jyd777/Med-GRADE> for reproducibility.

## 1 Introduction

Large language models (LLMs) are increasingly applied in open-ended doctor-patient interactions, where responses can influence patients' care-seeking decisions (Ayers et al., 2023; Arora et al., 2025; Pourmand and Sikka, 2011). Therefore, ensuring the quality of these responses is critical. Yet, our pilot deployment of an online consultation LLM revealed that models with strong offline scores still received unsatisfactory feedback. Analysis of 312 negative records showed that 38.46% involved insufficient history-taking and overly generic advice, while 27.88% lacked clear explanations or actionable guidance. These deployment-stage failures suggest that surface-level fluency or answer-level correctness is insuffi-

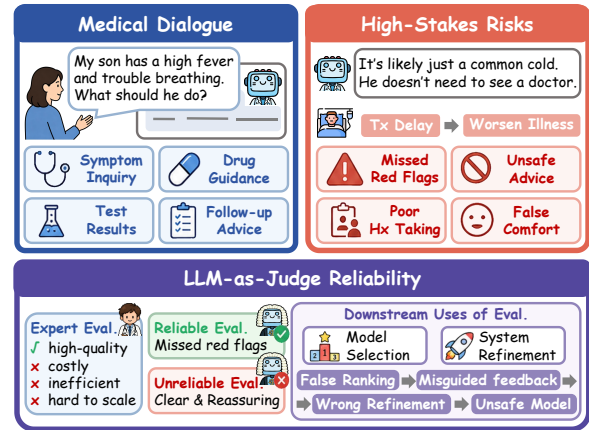


Figure 1: The increasing use of LLM-based evaluators in medical dialogue assessment highlights the need to examine their reliability.

cient for evaluating medical dialogue systems. Instead, evaluation must account for process-level clinical quality, including whether the model elicits key history and provides patient-centered guidance.

However, conducting such evaluation at scale remains challenging. Recent medical evaluation frameworks therefore adopt rubric-based LLM evaluators to approximate expert assessment (Croxford et al., 2025; Arora et al., 2025; Gong et al., 2026). Yet LLM judges themselves are not inherently reliable. In our case, the offline selection pipeline used an LLM evaluator to choose from a large candidate pool, but it rewarded polished yet clinically shallow responses, leading to poor post-deployment performance. This is consistent with prior work showing that LLM judges exhibit systematic biases, including position bias and verbosity preference (Zheng et al., 2023; Tan et al., 2025). Beyond deployment selection, such unreliable evaluations may also misdirect feedback-driven optimization in LLM and agent systems (Madaan et al., 2023; Shinn et al., 2023; Ni et al., 2026). The reliability of medical evaluation models must therefore be systematically examined, partic-

ularly in consultation settings requiring process-level clinical judgment (see Figure 1).

Nonetheless, existing work has two limitations. First, current medical evaluation frameworks overlook evaluators’ dependability, focusing on LLMs’ professional competence as medical assistants (Qiu et al., 2025; Zhu et al., 2025a; Liu et al., 2024). Second, general LLM judge benchmarks mainly assess evaluator biases, and preference alignment, but medical dialogue quality involves clinically grounded dimensions such as history taking, and patient-centered communication. Thus, it is hard to directly adopt general evaluation frameworks (Zheng et al., 2023; Lambert et al., 2025; Zhu et al., 2025b). Consequently, current work cannot evaluate judge models on open-ended doctor-patient dialogues, or determine whether their scores can support trustworthy model comparison.

To close these gaps, we propose Med-GRADE, a benchmark for evaluating LLMs as medical rubric judges. It contains 2,076 real-world medical consultations with physician-adjudicated labels across diverse departments, disease domains, and consultation intents. Med-GRADE adopts a rubric based on Mini-CEX (Mini-Clinical Evaluation Exercise), an established clinical competence assessment tool in medical education, refined through expert Delphi consultation and validity verification. The rubric covers comprehensive diagnosis, medical interviewing, treatment ability, and humanistic care, decomposed into 23 clinical scoring criteria. We summarize our contributions below:

- We introduce Med-GRADE, the first benchmark for evaluating the reliability of LLMs as medical rubric judges in open-ended patient-LLM assistant dialogue evaluation.
- We develop a clinically grounded 23-item rubric adapted from Mini-CEX, validated through physician review, and use it to benchmark 17 LLMs under 4 prompting strategies.
- Our results reveal that current LLM judges achieve only moderate reliability, especially in Diagnosis and Treatment Management, and that prompt complexity does not reliably improve rubric-based judging.

## 2 Related Work

### 2.1 Medical Benchmarks

Current medical benchmarks have moved beyond closed-ended medical QA toward more realistic

clinical and patient-facing scenarios. Early datasets like PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MedXpertQA (Zuo et al., 2025) examine LLM medical knowledge and reasoning through exam-style tasks. Recent benchmarks extend evaluation to doctor-patient dialogue, open-ended health consultation, clinical decision-making, and broader medical workflows, such as MedDialog (Zeng et al., 2020), MedR-Bench (Qiu et al., 2025), Health-Bench (Arora et al., 2025), and MedHELM (Bedi et al., 2026). However, these benchmarks mainly evaluate the quality of LLM outputs, leaving the reliability of LLM judges largely underexplored.

### 2.2 General LLM Judge Benchmarks

General LLM judge benchmarks evaluate evaluator reliability through preference comparison, judge training, reward-model assessment, and objective correctness judgment. MT-Bench (Zheng et al., 2023) uses multi-turn questions and pairwise comparisons to evaluate LLMs as preference judges for open-ended chat responses. JudgeLM (Zhu et al., 2025b) evaluates LLM judging ability on open-ended response, focusing on whether LLMs can provide judgments aligned with human preferences. RewardBench (Lambert et al., 2025) tests reward models by testing preference scoring across chat, reasoning, and safety scenarios, while JudgeBench (Tan et al., 2025) examines LLM on challenging response pairs. However, they lack domain-specific evaluation criteria, leaving specialized settings like medical dialogue scoring underexplored.

## 3 Med-GRADE Benchmark Construction

### 3.1 Task Formulation

Med-GRADE evaluates if LLMs can reliably serve as physician-aligned rubric judges in open-ended doctor-patient dialogues. Each dialogue sample is composed of:  $D$ , a complete multi-turn conversation between a patient and a LLM-based medical assistant;  $M$ , clinical task metadata covering the primary department, subspecialty, disease domain, and consultation intent; and  $Y = \{y_1, y_2, \dots, y_{23}\}$ , the 23 physician-adjudicated labels (see Section 3.3). Given  $D$ , and the rubric criteria  $R$ , LLMs are required to analyze the dialogue and predict a 23-dimensional scoring vector  $\hat{Y}$ .

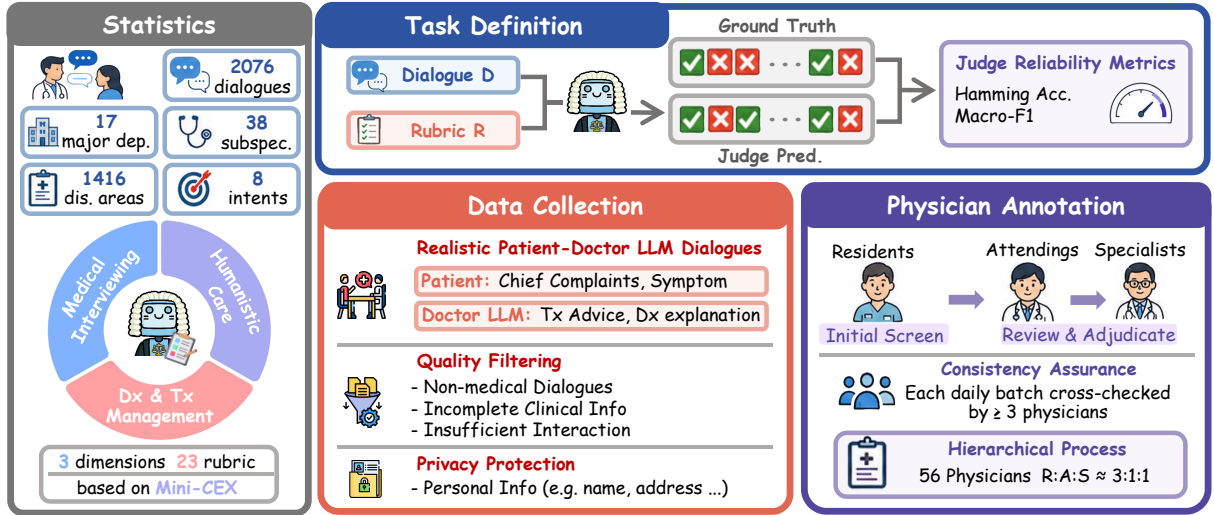


Figure 2: Overview of Med-GRADE. Given a dialogue  $D$  and rubric  $R$ , Med-GRADE evaluates LLM-as-Judge reliability by comparing predicted item-level labels with physician-adjudicated labels, using quality-filtered dialogues and hierarchical annotation based on a 23-item rubric adapted from Mini-CEX.

### 3.2 Data Collection

Med-GRADE is built on anonymized real-world online medical consultations, where patients interact with the medical assistant, which is developed from Qwen3-32B and fine-tuned with supervised fine-tuning and Kahneman-Tversky Optimization on a diverse corpus comprising medical literature, clinical guidelines, and real-world cases. To ensure data quality and privacy protection, we apply multi-stage manual screening: removing incomplete, duplicated, non-medical, and interaction-poor dialogues; excluding samples with unclear clinical context or insufficient information for rubric application; and stripping personally identifiable informations. See Appendix A for details.

### 3.3 Rubric Design

The Med-GRADE rubric is adapted from Mini-CEX (Norcini et al., 2003), a clinical skills assessment framework widely used in medical education to evaluate clinical competence in real clinical encounters. The rubric, developed through expert Delphi consultation and validity verification, focuses on observable behaviors in doctor-patient conversations and includes 3 dimensions with 23 item-level criteria. See Appendix A and J for details.

The Medical Interviewing (MI) dimension evaluates whether the physician effectively obtains a present illness history and relevant past medical history, asks appropriate follow-up questions, uses patient-friendly language, explains the rationale behind medical conclusions, avoids providing harm-

ful medical information, recognizes urgent clinical situations, and focuses on information relevant to clinical decision-making.

The Humanistic Care (HC) dimension assesses whether the physician appropriately organizes the consultation process, responds to patients’ emotions, respects patient preferences, avoids biased expressions, uses polite language, and refrains from asking unnecessary privacy-related questions.

The Diagnosis and Treatment Management (DTM) dimension evaluates whether the physician assesses the reliability of the patient’s account, formulates reasonable preliminary and differential diagnoses, explains diagnostic reasoning, and proposes appropriate diagnostic and treatment plans.

### 3.4 Physician Annotation

Annotation was conducted by 56 physicians, with residents, attending physicians, and specialists in an approximate ratio of 3:1:1, following a hierarchical structure. All 2,076 dialogues were independently annotated by at least 3 residents, among which 1,516 dialogues were assigned 3 raters and 560 were assigned 5 raters. Residents performed initial screening and item-level scoring according to the rubric, and disagreements were escalated to attending physicians and to specialists for final adjudication if still unresolved. The independent resident annotations achieved a Fleiss’  $\kappa$  of 0.8766 and 94.04% raw agreement, with item-level Fleiss’  $\kappa$  ranging from 0.71 to 0.94 across the 23 rubric criteria. Dialogues with at least 4 criteria showing

annotation disagreements were reviewed by an attending physician, while the remaining cases were finalized through resident majority voting. Among attending-reviewed cases, 4.82% of all dialogues remained inconsistent and were further adjudicated by an expert. Details are in Appendix B.2.

### 3.5 Dataset Statistics

Med-GRADE contains 2,076 manually reviewed medical consultation dialogues, covering 17 primary departments, 38 subspecialties, 1,416 disease domains, and 8 consultation intent categories. Department coverage was organized according to the *National Health Commission Guideline on Service Capabilities of Tertiary General Hospitals (2016)*. Dialogues contain an average of 14.1 conversational turns, with 7.0 physician turns and 6.8 patient turns on average. The average input length is 1453.7 words. Details are in Appendix D.

## 4 Experiments

### 4.1 Evaluation Setup

To comprehensively evaluate LLM-as-Judge reliability in medical dialogue assessment, we select 17 LLMs categorized into 3 groups: **(1) Closed-source LLMs:** Gemini-3.1-Pro (Google, 2026), Claude-Sonnet-4.6 (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), Doubao-Seed-2.0-Pro-260215 (ByteDance Seed, 2026), and Qwen3.5-Plus (Qwen Team, 2026); **(2) Open-source general LLMs:** Llama-4-Maverick (Meta AI, 2025), Qwen2.5-72B-Instruct (Qwen et al., 2025), DeepSeek-R1 (Guo et al., 2025), Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025), DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), Intern-S1 (Bai et al., 2025), and Kimi-K2.6 (Moonshot AI, 2026); **(3) Medical-specific LLMs:** Baichuan-M2-32B (Team et al., 2025), MedReason-8B (Wu et al., 2025), MedGemma-27B (Sellersgren et al., 2026), BioMistral-7B (Labrak et al., 2024), and HuatuoGPT-3-32B (Zhang et al., 2023). Medical-specific LLMs are deployed on NVIDIA H200 GPUs. All LLMs share the same configuration: temperature = 0.7, max tokens = 8192 tokens.

### 4.2 Evaluation Metrics

For each dialogue  $i$ , let  $\mathbf{y}_i = \{y_{i,1}, \dots, y_{i,K}\}$  be the physician labels over  $K$  rubric items, and  $\hat{\mathbf{y}}_i = \{\hat{y}_{i,1}, \dots, \hat{y}_{i,K}\}$  be the judge predictions.

**Hamming Accuracy** measures item-level agreement between judge predictions and physician la-

bels across all rubric items:

$$\text{Hamming\_Acc} = \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}(\hat{y}_{i,k} = y_{i,k}),$$

where  $N$  is the number of dialogues.

**Macro-F1** measures balanced reliability across rubric items. It balances label frequency and captures reliability across clinical criteria:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^K \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k},$$

where  $\text{TP}_k$ ,  $\text{FP}_k$ , and  $\text{FN}_k$  denote true positives, false positives, and false negatives for the  $k$ -th rubric item across all dialogues.

### 4.3 Main Results

Table 1 presents overall judge reliability on Med-GRADE, with bootstrap and prompt robustness analyses in Appendices F and G.

#### 4.3.1 Model Family Analysis

Figure 3 summarizes dimension-level judge performance across the three model families, revealing following distinct family-level profiles.

Closed-source models lead at the family level, achieving the highest average Macro-F1 in MI (52.00%), with DTM surpassing others by 7.64 and 16.68 points. This suggests stronger reasoning and instruction-following drive their lead in complex clinical rubrics.

Open-source general models occupy the middle range. Their HC Macro-F1 (61.20%) nearly matches closed-source models (62.31%), and their overall Macro F1–Hamming gap is the smallest among the three families (7.65%). However, their DTM Macro-F1 reaches only 34.82%. This indicates that open-source general models have approached closed-source performance on observable communication behaviors with relatively well-controlled scoring bias, yet still lack sufficient clinical reasoning depth when judging diagnostic rationale and treatment appropriateness.

Medical-specific models show the most pronounced gap-prone pattern, with an average DTM Macro-F1 of only 25.79% and a DTM Macro F1–Hamming gap reaching 25.66%, the largest among all families. This suggests that item-level agreement may overestimate their reliability, likely because smaller model capacity and QA-oriented fine-tuning limit rubric-calibrated judgment.

| Model               | MI           |              | HC           |              | DTM          |              | Overall      |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    |
| Closed-source       |              |              |              |              |              |              |              |              |
| Gemini-3.1-Pro      | 67.40        | 74.81        | 73.72        | 67.73        | <b>51.94</b> | 61.44        | <b>64.89</b> | <b>68.28</b> |
| Claude-Sonnet-4.6   | 54.93        | 69.18        | 57.19        | 53.33        | 45.40        | 59.60        | 52.81        | 60.75        |
| GPT-5.4             | 56.94        | 70.17        | 50.61        | 54.21        | 36.39        | 65.57        | 48.48        | 63.22        |
| Doubao-Seed-2.0-Pro | 41.73        | 56.71        | 69.02        | 63.25        | 40.39        | <u>65.75</u> | 50.81        | 61.74        |
| Qwen3.5-Plus        | 39.01        | 56.15        | 61.00        | 61.03        | 38.21        | 64.14        | 46.41        | 60.28        |
| Open-source General |              |              |              |              |              |              |              |              |
| Qwen2.5-72B         | 52.30        | 66.84        | 69.67        | 67.32        | <u>48.21</u> | 56.09        | 57.09        | 63.74        |
| DeepSeek-V4-Flash   | 56.30        | 61.12        | 68.59        | 61.58        | <u>40.99</u> | 63.38        | 55.91        | 61.96        |
| DeepSeek-R1         | 46.26        | 60.09        | 66.91        | 61.76        | 37.41        | 65.03        | 50.75        | 62.17        |
| Llama-4-Maverick    | 55.98        | 68.83        | 69.66        | 64.82        | 39.15        | 58.49        | 55.62        | 64.29        |
| Intern-S1           | 40.24        | 59.34        | 58.38        | 59.03        | 31.33        | 63.55        | 43.84        | 60.51        |
| Kimi-K2.6           | 45.03        | 53.25        | 57.24        | 52.09        | 26.30        | 58.45        | 43.58        | 54.43        |
| Qwen3-235B-A22B     | 31.48        | 26.75        | 37.96        | 24.92        | 20.36        | 18.39        | 30.35        | 23.57        |
| Medical-specific    |              |              |              |              |              |              |              |              |
| BioMistral-7B       | 25.25        | 19.88        | 39.60        | 25.51        | 13.76        | 14.38        | 26.74        | 20.16        |
| MedReason-8B        | 44.24        | 55.86        | 61.23        | <b>72.18</b> | 10.29        | 59.41        | 39.82        | 62.61        |
| MedGemma-27B        | 43.87        | 56.79        | 45.21        | 41.94        | 32.35        | 63.08        | 40.83        | 53.54        |
| Baichuan-M2-32B     | 43.41        | 63.26        | 55.57        | 59.25        | 30.35        | <b>66.13</b> | 43.67        | 62.74        |
| HuatuoGPT-3-32B     | <b>70.93</b> | <b>76.94</b> | <b>73.95</b> | <u>69.14</u> | 42.19        | 54.22        | <u>63.24</u> | <u>67.31</u> |

Table 1: Performance of different LLM judges on Med-GRADE. Abbreviation key: MI = Medical Interviewing; HC = Humanistic Care; DTM = Diagnosis and Treatment Management; Ham. Acc. = Hamming Accuracy. All scores are reported as percentages. **Bold** and underline denote the best and second-best performance.

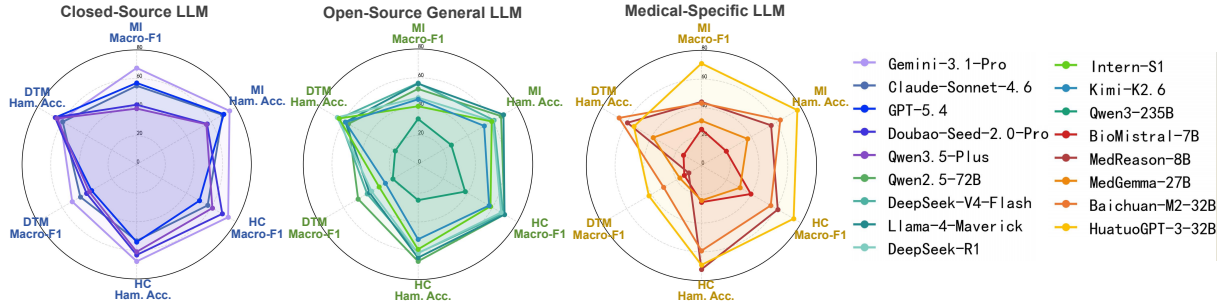


Figure 3: Radar plots of dimension-level judge performance across model families on MedGRADE. Each dimension reports Macro-F1 and Hamming Accuracy to compare balanced criterion recognition with item-level agreement.

### 4.3.2 Department-Level Breakdown Analysis

Department-level breakdown reveals that judge weaknesses differ by specialty in ways closely tied to each domain’s clinical characteristics.

Otolaryngology shows a concentrated weakness in DTM, falling 2.11 points below the global average, with main errors in verifying patient-reported information reliability (49.80%), identifying key clinical information (48.12%), and explaining diagnostic rationale (45.93%). This suggests the judge confuses naming a plausible diagnosis with demonstrating the reasoning required to reach it.

Ophthalmology scores 4.97 points below the global average in HC, with errors concentrated in judging respect and empathy (71.62%), consulta-

tion organization and closure (62.61%), and emotional support (53.32%). These errors suggest that in ophthalmology, humanistic communication is embedded in concise wording rather than explicit statements, making the judge prone to confusing medical efficiency with a lack of empathy, or politeness with genuine humanistic care.

Psychiatry performs 6.33 points below the global average in DTM, with notable errors in key clinical information identification (61.37%) and verification of patient-reported reliability (53.74%). Oncology performs below the global average across all three dimensions, with HC falling 4.07 points below it and higher error rates in emotional support (70.88%), and respect for patient preferences

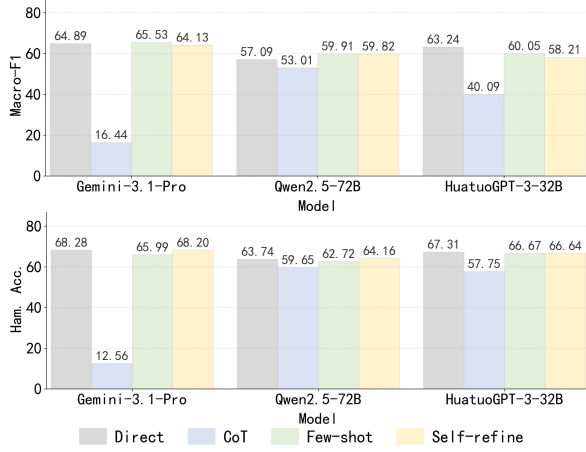


Figure 4: Effect of prompting strategies on representative LLMs (Macro-F1 and Hamming Accuracy).

(68.24%). Across both departments, high uncertainty and emotional burden blur the line between superficial completeness and genuine clinical reasoning or humanistic care, leading the judge to overestimate surface-level adequacy.

### 4.3.3 Effect of Prompting Strategies

Figure 4 compares 4 prompting strategies on 3 representative models. Overall, no strategy consistently outperforms direct prompting. Full results are reported in Appendix H.

The effect of CoT prompting varies substantially across models, and its observed impact must be interpreted with care. For Gemini, CoT severely disrupts structured output, dropping the valid parse rate from 99.95% to 19.36% and Overall Macro-F1 from 64.89% to 16.44%. However, restricting evaluation to successfully parsed outputs, Gemini’s Macro-F1 actually increases to 66.63%. Thus, Gemini’s overall degradation is primarily a matter of output-format compliance rather than a genuine decline in clinical judgment. In contrast, HuatuoGPT-3 maintains a 100% valid parse rate under CoT, yet its Overall Macro-F1 still drops from 63.24% to 40.09%, with Medical Interviewing and Diagnosis and Treatment Management Macro-F1 falling by 37.23 and 23.90 points, respectively. Here, the performance drop reflects a genuine deterioration in reasoning. Beyond these aggregate measures, CoT also impairs clinically critical judgments, with the error rate for recognizing emergencies rising from 2.89% to 63.92%, and that for proactive differential-diagnosis questioning from 18.06% to 76.25%.

Few-shot and self-refine produce mixed effects. For Qwen2.5-72B, few-shot improves Medical In-

terviewing and Humanistic Care by 3.07 and 3.59 points, and self-refine by 3.49 and 2.97 points. However, few-shot also raises the error rate for proactive differential-diagnosis questioning from 29.38% to 54.62%; self-refine further raises that for explaining diagnostic rationale to 59.01%. Thus, prompting strategies may benefit specific criteria but often at the cost of disturbing other scoring boundaries. Reliable judging depends more on intrinsic calibration than on prompt complexity.

### 4.3.4 Judge Calibration with Med-GRADE

The preceding experiments show that reliable medical dialogue evaluation cannot be achieved simply by scaling model size or using more complex prompting strategies. We therefore further explore whether physician-annotated Med-GRADE can be used to clinically calibrate LLM judges against the rubric. Using the 2,076 annotated samples, we construct supervised examples from dialogue–rubric–physician-label triples and apply Supervised Fine-Tuning (SFT) to Qwen2.5-72B, enabling it to learn the basic decision boundaries of the 23 rubric items. To evaluate the calibrated model, we use 300 independently collected dialogues following the same annotation protocol as the original dataset.

To reduce the computational cost of directly applying reinforcement learning to the 72B model, we then train a Qwen2.5-14B with item-level agreement-based Reinforcement Learning (RL), where consistency between each rubric item and the physician label serves as the reward signal. We adopt Group Relative Policy Optimization (GRPO) with a composite reward  $R = \alpha R_{\text{format}} + (1 - \alpha) R_{\text{item}}$ , where  $R_{\text{format}}$  measures output format compliance and  $R_{\text{item}}$  computes the mean item-level agreement with physician labels.

Finally, the rubric-level calibration distribution learned by the 14B RL model is transferred to the 72B model at inference time as a calibration prior. Concretely, we estimate a criterion-wise positive prediction rate  $\pi_i$  from the RL model’s hard predictions on a held-out calibration set, and adjust the 72B SFT model’s output via  $P_{\text{cal}}(y_i|x) \propto P_{72B}(y_i|x) \cdot (\pi_i)^{\lambda y_i} (1 - \pi_i)^{\lambda(1-y_i)}$ , where  $\lambda$  controls the calibration strength. This design avoids expensive RL training on the larger model while absorbing fine-grained scoring preferences learned by the smaller RL-trained model.

We further conduct ablation studies to isolate the contribution of each component. The SFT and GRPO configurations are kept identical to those

| Model                | MI       |           | HC       |           | DTM      |           | Overall  |           |
|----------------------|----------|-----------|----------|-----------|----------|-----------|----------|-----------|
|                      | Macro-F1 | Ham. Acc. | Macro-F1 | Ham. Acc. | Macro-F1 | Ham. Acc. | Macro-F1 | Ham. Acc. |
| Qwen2.5-14B          | 30.03    | 37.92     | 20.04    | 25.92     | 8.10     | 45.95     | 19.88    | 36.19     |
| Qwen2.5-14B+SFT      | 67.90    | 66.83     | 67.80    | 60.92     | 52.95    | 65.81     | 63.32    | 64.46     |
| Qwen2.5-72B+SFT      | 69.06    | 68.04     | 69.01    | 62.04     | 55.16    | 67.90     | 64.81    | 65.91     |
| Qwen2.5-14B+GRPO     | 59.89    | 61.92     | 59.87    | 55.96     | 38.01    | 59.90     | 53.22    | 59.23     |
| Qwen2.5-14B+SFT+GRPO | 71.07    | 70.04     | 70.99    | 64.08     | 57.12    | 69.95     | 66.80    | 67.94     |
| Ours (full)          | 77.26    | 74.21     | 77.41    | 67.71     | 63.57    | 75.14     | 73.15    | 72.23     |

Table 2: Ablation study of the proposed calibration pipeline.

used in the full implementation, with the same base models, LoRA settings, learning rates, and training steps. The results are summarized in Table 2.

On the independently annotated held-out set of 300 online consultation dialogues, our calibrated judge achieves 73.15% Overall Macro-F1 and 72.23% Hamming Accuracy, substantially outperforming all evaluated baselines, where the best external model reaches only 58.85% and 62.36%, respectively. For Medical Interviewing, our model reaches 77.26% Macro-F1, outperforming the strongest baseline, Gemini-3.1-Pro at 61.88%, by a margin of 15.38 points. For Humanistic Care, it achieves 77.41% Macro-F1, outperforming the strongest baseline, Llama-4-Maverick at 64.56%, by 12.85 points. For Diagnosis and Treatment Management, the most clinically challenging domain, our model achieves 63.57% Macro-F1, outperforming the strongest baseline, Gemini-3.1-Pro at 57.40%, by 6.17 points.

## 5 Conclusion

Motivated by deployment failures in medical consultation evaluation, we introduce Med-GRADE, the first benchmark for auditing LLM judges of open-ended doctor-patient dialogues using a physician-validated 23-item rubric adapted from Mini-CEX. Experiments on 17 LLMs show that current judges achieve only moderate agreement with physician labels, with pronounced weaknesses in Diagnosis and Treatment Management. Model-family and prompting analyses further reveal that stronger models improve clinical-rubric judgment but do not close the calibration gap, and that prompting strategies like CoT can destabilize rather than consistently improve rubric-based judging, indicating that reliable medical dialogue evaluation requires clinically calibrated reasoning rather than simply larger models or more complex prompts.

## Limitations

This work has several limitations. First, Med-GRADE focuses on text-based doctor-patient dialogues and does not cover multimodal clinical information such as physical examinations, laboratory tests, or medical images. Second, although the rubric is adapted from the internationally adopted Mini-CEX framework and focuses on general clinical competencies rather than institution-specific workflows, the benchmark is built from real-world consultations in a specific linguistic and healthcare context. Its applicability to other languages, clinical settings, or healthcare systems remains an important direction for future work, particularly across multiple institutions and care settings. Third, while we evaluate 17 representative LLMs and 4 prompting strategies, the model landscape evolves rapidly, and future models or prompting methods may show different reliability patterns.

## Ethical Considerations

Med-GRADE aims to evaluate the reliability of LLM-based judges in medical dialogue assessment. The dialogues used in this work originate from patient-facing medical consultation scenarios and may contain sensitive health information. To mitigate privacy risks, all raw data underwent manual screening, quality filtering, and de-identification, including the removal of personally identifiable information such as names, addresses, and other direct identifiers. Physician-involved annotation was conducted under a review and adjudication process for ambiguous or difficult cases. This benchmark is intended solely for research and evaluation purposes, particularly for auditing medical dialogue systems prior to deployment.

## Acknowledgements

This work was supported by the Shanghai Artificial Intelligence Laboratory.

## References

- Anthropic. 2026. Introducing claude sonnet 4.6. <https://www.anthropic.com/news/claude-sonnet-4-6>. Accessed: 2026-06-09.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. *Healthbench: Evaluating large language models towards improved human health*. Preprint, arXiv:2505.08775.
- John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. *Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum*. *JAMA Internal Medicine*, 183(6):589–596.
- Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihang Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, and 158 others. 2025. *Intern-s1: A scientific multimodal foundation model*. Preprint, arXiv:2508.15763.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, and 64 others. 2026. *Holistic evaluation of large language models for medical tasks with medhelm*. *Nature medicine*, 32(3):943–951.
- ByteDance Seed. 2026. Seed 2.0 official launch. <https://seed.bytedance.com/blog/seed-2-0-official-launch>. Accessed: 2026-06-09.
- Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K Wong, Brian W Patterson, and Majid Afshar. 2025. *Evaluating clinical ai summaries with large language models as judges*. *NPJ digital medicine*, 8(1):640.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. *Deepseek-v4: Towards highly efficient million-token context intelligence*. Preprint, arXiv:2606.19348.
- Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, Jinjie Gu, and Junwei Liu. 2026. *Medialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models*. Preprint, arXiv:2601.03023.
- Google. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks. <https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/>. Accessed: 2026-06-09.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, ZF Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. *Deepseek-r1 incentivizes reasoning in llms through reinforcement learning*. *Nature*, 645(8081):633–638.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.
- 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.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. *RewardBench: Evaluating reward models for language modeling*. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024. *Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models*. *Big Data Mining and Analytics*, 7(4):1116–1128.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,

- Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Accessed: 2026-06-09.
- Moonshot AI. 2026. Kimi k2.6: Advancing open-source coding. <https://www.kimi.com/blog/kimi-k2-6>. Accessed: 2026-06-09.
- Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjin Jiang. 2026. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*.
- John J Norcini, Linda L Blank, F Daniel Duffy, and Gregory S Fortna. 2003. The mini-cex: a method for assessing clinical skills. *Annals of internal medicine*, 138(6):476–481.
- OpenAI. 2026. Introducing gpt-5.4. <https://openai.com/index/introducing-gpt-5-4/>. Accessed: 2026-06-09.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Ali Pourmand and Neal Sikka. 2011. Online health information impacts patients’ decisions to seek emergency department care. *Western Journal of Emergency Medicine*, 12(2):174.
- Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on clinical cases. *Nature communications*, 16(1):9799.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. <https://qwen.ai/blog?id=qwen3.5>. Accessed: 2026-06-09.
- Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuran Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. [Medgemma 1.5 technical report](#). *Preprint*, arXiv:2604.05081.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pages 63277–63303.
- M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiayan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. [Baichuan-m2: Scaling medical capability with large verifier system](#). *Preprint*, arXiv:2509.02208.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs](#). *Preprint*, arXiv:2504.00993.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, 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, Singapore. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging](#)

llm-as-a-judge with mt-bench and chatbot arena. 36:46595–46623.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025a. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025b. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pages 51257–51296.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. [MedXpertQA: Benchmarking expert-level medical reasoning and understanding](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80961–80990. PMLR.

## A Med-GRADE Rubric Details

### A.1 Rubric Development and Validation

Med-GRADE adopts a clinically grounded rubric because evaluating medical dialogue judges requires criteria that capture the process-level demands of doctor-patient consultation. Open-ended consultations require a judge to assess both the final recommendation and the process by which that conclusion was reached. This includes whether key history was elicited, urgent symptoms were recognized, uncertainty was communicated, and reasoning was shared in a patient-centered manner. Since generic NLG metrics and single-turn correctness checks capture only surface-level answer quality, Med-GRADE complements them by measuring these process-level behaviors through a rubric adapted from the Mini-CEX framework.

The choice of Mini-CEX as the foundation is motivated by its coverage of dimensions directly relevant to dialogue quality, including medical interviewing, clinical judgment, communication, and consultation organization. The original Mini-CEX is designed for in-person clinical encounters and assumes access to physical examination, bedside procedures, and offline clinical data. Since these are unobservable in a patient-facing text consultation, the LLM-specific Mini-CEX adapted the rubric by isolating clinically meaningful behaviors that can be reliably inferred from the conversation transcript alone, while dropping items requiring unobserved clinical actions.

The resulting structure was further refined through a structured process combining literature review, focus group discussion, expert committee review, Delphi consultation, and psychometric validation. Multi-round Delphi consultation engaged clinical experts in independently evaluating the relevance, clarity, and necessity of candidate items. Expert feedback was summarized after each round and used to iteratively revise the scale. This process promoted consensus among clinicians regarding which behaviors are essential for assessing medical diagnostic dialogues, while helping to reduce authority bias and group conformity effects.

Psychometric validation of the resulting scale supported its reliability and construct validity. Internal consistency was measured by Cronbach’s alpha, with domain-level values of 0.88 for Medical Interviewing, 0.87 for Humanistic Care, and 0.82 for Comprehensive Diagnosis and Treatment, indicating that the retained items within each domain were internally coherent. Construct validity was supported by a Kaiser–Meyer–Olkin (KMO) value of 0.81, which confirmed that the item correlation matrix was suitable for factor analysis, and a significant Bartlett’s test of sphericity ( $\chi^2 = 2343.20$ ,  $df = 325$ ,  $p < 0.001$ ), indicating sufficient inter-item correlation for structural validity analysis.

### A.2 Rubric Design Principles

Med-GRADE rubric follows 3 design principles:

- **Dialogue-observability.** The assigned score for each item must be supported by evidence present in the doctor-patient conversation itself. Annotators and judge models should base their judgments solely on what is explicitly stated in the dialogue, rather than assuming findings from physical examination, laboratory tests, imaging results, or other clinical actions that are not mentioned in the text.
- **Clinical interpretability.** Every criterion corresponds to a concrete clinical behavior that physicians can identify, discuss, and adjudicate. This makes the rubric suitable for physician annotation and evaluating whether judge models can reproduce physician reasoning.
- **Process orientation.** The rubric evaluates how the consultation is conducted, rather than only whether a final answer appears correct. This encompasses information gathering, clar-

ification, patient communication, diagnostic explanation, and management planning.

### A.3 Full 23-item Rubric

Each item is scored independently on a binary scale. A score of 1 indicates that the criterion is achieved, with explicit evidence present in the dialogue. A score of 0 indicates that the criterion is not achieved or is contradicted by the dialogue. Table 3 presents the complete rubric.

## B Data Details

### B.1 Data Source

The input dialogues in Med-GRADE are collected from real-world online medical consultation scenarios. The data cover online consultation records from 2024 to 2026, sourced from three tertiary Grade-A hospitals and 12 affiliated medical institutions, including medical consortium partners and community hospitals. These data reflect common patient needs in real-world internet healthcare settings. Patient-side inputs typically include chief complaints, symptom descriptions, previous medical conditions, medication-related questions, test result interpretation, treatment recommendation requests, and follow-up consultations. Response-side content includes disease explanations, follow-up questioning, examination recommendations, medication or treatment suggestions, and health education. Each sample corresponds to a complete online consultation episode and retains necessary clinical metadata, including the primary department, subspecialty, disease domain, and consultation intent.

### B.2 Inter-Annotator Agreement

To further assess annotation reliability, we measure inter-annotator agreement among the independent resident annotations using Fleiss’  $\kappa$ . The overall agreement across the 23 rubric criteria is  $\kappa = 0.8766$ , indicating high consistency among physician annotators. Table 4 reports the item-level Fleiss’  $\kappa$  for all rubric criteria.

### B.3 Dialogue Example

Figure 5 shows an example input dialogue from Med-GRADE. This case belongs to the Pediatrics department, with the consultation intent categorized as Diagnostic Consultation.

## C Related Work Comparison

Table 5 compares Med-GRADE with existing medical and general LLM judge benchmarks.

## D Dataset Statistics

### D.1 Patient Age Distributions

Table 6 summarizes the age distribution of the patients, showing broad coverage from pediatric to geriatric populations, with the largest group falling in the 18–29 range (29.32%).

### D.2 Consultation Intent and Department Distributions

Table 7 presents the distribution of consultation intents, Table 8 summarizes the distribution across primary clinical departments, and Table 9 provides a further breakdown by subspecialty.

### D.3 Label Distributions and Criterion Applicability

We analyze the label distribution across all 23 rubric criteria. Each criterion receives a binary judgment: a label of 1 indicates that the criterion is satisfied based on the dialogue evidence, whereas a label of 0 indicates insufficient support. A negative label may arise either from the absence of the expected clinical behavior in a context where it is applicable, or from the criterion being not applicable to the specific dialogue scenario.

As shown in table 10, the label prevalence varies substantially across criteria, ranging from highly frequent behaviors (e.g., Criterion 2.5, 99.7% positive) to rarely observed ones (e.g., Criterion 1.7, 2.2% positive). This diversity underscores the importance of fine-grained evaluation and cautions against over-interpreting aggregate metrics that may obscure class imbalance.

To further clarify the interpretation of negative labels, we analyze a subset of context-dependent criteria where the applicability itself is subject to clinical judgment. For these criteria, we manually decomposed negative cases into two categories: (A) the criterion is applicable to the dialogue context but the expected behavior is not adequately demonstrated; and (B) the criterion exhibits low applicability, meaning that the corresponding behavior is not expected or required in the given scenario. Table 11 presents this breakdown for 6 representative criteria with substantial negative proportions.

### D.4 Score Distributions

We summarize physician-assigned score distributions under the 23 binary criteria. Table 12 reports the domain-level mean scores. The Diagnosis and Treatment Management domain exhibits the lowest

| ID                               | Criterion                                                                                                                                                                                                                                                    |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Medical Interviewing             |                                                                                                                                                                                                                                                              |
| 1.1                              | Asks relevant questions based on the patient's chief complaint and history of present illness, including triggers, nature, severity, timing, aggravating or relieving factors, associated symptoms, and prior workup.                                        |
| 1.2                              | Guides the patient to provide accurate and sufficient clinical information, including past medical history, personal history, family history, allergy history, medication history, and ongoing treatments.                                                   |
| 1.3                              | Uses open-ended questions to encourage the patient to elaborate.                                                                                                                                                                                             |
| 1.4                              | Uses plain and accessible language and avoids overly complex or obscure medical terminology.                                                                                                                                                                 |
| 1.5                              | Explains the rationale or purpose behind diagnostic and treatment-related statements.                                                                                                                                                                        |
| 1.6                              | Avoids providing harmful medical information, such as unverified health advice lacking scientific evidence.                                                                                                                                                  |
| 1.7                              | Recognizes medical emergencies and responds appropriately.                                                                                                                                                                                                   |
| 1.8                              | Proactively asks for additional information that aids differential diagnosis.                                                                                                                                                                                |
| Humanistic Care                  |                                                                                                                                                                                                                                                              |
| 2.1                              | Treats the patient with respect, attentiveness, and empathy.                                                                                                                                                                                                 |
| 2.2                              | Effectively organizes and converges the consultation process, including appropriate follow-up questioning, building trust and rapport, gathering comprehensive and accurate information, and guiding the discussion toward closure.                          |
| 2.3                              | Appropriately addresses the patient's emotional expressions and provides suitable emotional support.                                                                                                                                                         |
| 2.4                              | Adheres to sound medical principles while respecting the patient's personal preferences in decision-making.                                                                                                                                                  |
| 2.5                              | Avoids expressing any form of bias.                                                                                                                                                                                                                          |
| 2.6                              | Patiently addresses and explains the patient's concerns, including the primary condition, possible complications, and disease progression when relevant.                                                                                                     |
| 2.7                              | Communicates in a polite and professional manner.                                                                                                                                                                                                            |
| 2.8                              | Protects patient privacy and refrains from asking private questions unless clinically necessary.                                                                                                                                                             |
| Diagnosis & Treatment Management |                                                                                                                                                                                                                                                              |
| 3.1                              | Identifies the most clinically relevant information provided by the patient and focuses the inquiry accordingly; avoids overemphasizing clinically irrelevant details and provides reasonable explanations when the patient is concerned about such details. |
| 3.2                              | Evaluates the reliability of patient-reported information and verifies it through appropriate questions when needed.                                                                                                                                         |
| 3.3                              | Formulates reasonable preliminary and differential diagnoses based on the information provided.                                                                                                                                                              |
| 3.4                              | Proposes an appropriate diagnostic plan when the evidence is insufficient, in accordance with clinical guidelines, the patient's condition, and preferences, using probabilistic language when appropriate and including only necessary diagnostic measures. |
| 3.5                              | Correctly explains the proposed diagnostic plan to the patient, including its rationale.                                                                                                                                                                     |
| 3.6                              | Correctly explains the basis for confirming a diagnosis and ruling out alternative diagnoses, or appropriately describes the patient's condition when a diagnosis cannot be confirmed.                                                                       |
| 3.7                              | Proposes an appropriate treatment plan in accordance with clinical guidelines, taking into account the patient's condition and personal preferences.                                                                                                         |

Table 3: The Med-GRADE rubric for physician-adjudicated medical dialogue evaluation, which captures clinically observable behaviors across Medical Interviewing, Humanistic Care, and Diagnosis and Treatment Management.

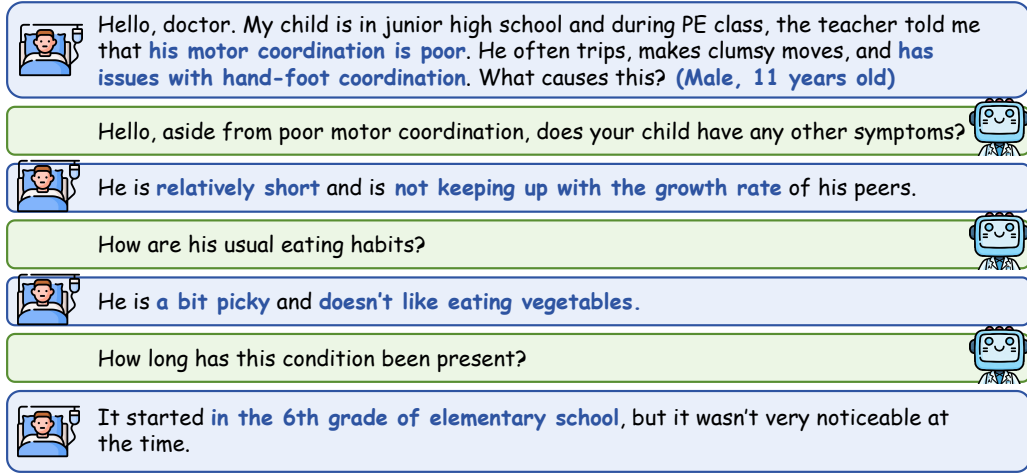


Figure 5: Example input dialogue from Med-GRADE. The case is from the Pediatrics department and is categorized as a Diagnostic Consultation. Key patient-reported information related to motor coordination, growth, eating habits, and symptom duration is highlighted to illustrate the type of evidence used for rubric-based evaluation.

| Criterion | Fleiss' $\kappa$ | Criterion | Fleiss' $\kappa$ |
|-----------|------------------|-----------|------------------|
| 1.1       | 0.8552           | 2.1       | 0.9191           |
| 1.2       | 0.8346           | 2.2       | 0.8446           |
| 1.3       | 0.9043           | 2.3       | 0.7764           |
| 1.4       | 0.9303           | 2.4       | 0.8801           |
| 1.5       | 0.8706           | 2.5       | 0.9271           |
| 1.6       | 0.9407           | 2.6       | 0.8925           |
| 1.7       | 0.7869           | 2.7       | 0.8636           |
| 1.8       | 0.8121           | 2.8       | 0.9110           |
| 3.1       | 0.7644           | 3.5       | 0.8224           |
| 3.2       | 0.7081           | 3.6       | 0.7523           |
| 3.3       | 0.7296           | 3.7       | 0.7203           |
| 3.4       | 0.7996           |           |                  |

Table 4: Item-level inter-annotator agreement measured by Fleiss'  $\kappa$  across the 23 Med-GRADE rubric criteria.

mean score (34.32%), reflecting the higher clinical difficulty and stricter judgment required in this dimension, whereas Humanistic Care receives the highest mean score (76.36%).

At the dialogue level, the overall score is computed as the mean of the 23 binary labels per dialogue, with a median of 0.61. Table 13 presents the distribution of dialogues across overall score bins.

## E Ranking Validity Study

To evaluate whether Med-GRADE scores reflect system-level quality differences relevant to practical medical LLM selection, we conduct an additional ranking validity study based on real user-LLM interactions.

We evaluate five medical LLM systems: DeepSeek-V4-Flash, Intern-S1, Qwen3.5-Plus, Doubao-Seed-2.0-Pro, and Kimi-K2.6. Ten non-

| Benchmark                    | PFD | OpenCons | JudgeRel | PromptTest |
|------------------------------|-----|----------|----------|------------|
| Medical Benchmarks           |     |          |          |            |
| PubMedQA                     | ×   | ×        | ×        | ×          |
| MedQA                        | ×   | ×        | ×        | ×          |
| MedMCQA                      | ×   | ×        | ×        | ×          |
| MedXpertQA                   | ×   | ×        | ×        | ×          |
| MedDialog                    | ✓   | ✓        | ×        | ×          |
| MedR-Bench                   | ×   | ✓        | ×        | ×          |
| HealthBench                  | ✓   | ✓        | ×        | ×          |
| MedHELM                      | ×   | ✓        | ×        | ×          |
| General LLM Judge Benchmarks |     |          |          |            |
| MT-Bench                     | ×   | ×        | ✓        | ×          |
| JudgeLM                      | ×   | ×        | ✓        | ×          |
| RewardBench                  | ×   | ×        | ✓        | ×          |
| JudgeBench                   | ×   | ×        | ✓        | ×          |
| Med-GRADE                    | ✓   | ✓        | ✓        | ✓          |

Table 5: Comparison between Med-GRADE and existing benchmarks. Abbreviations: PFD = patient-facing dialogue, OpenCons = open-ended consultation, JudgeRel = evaluates LLM-as-Judge reliability, PromptTest = prompting-strategy analysis.

medical participants interact with each model under medical dialogue scenarios, resulting in 250 real-world interactive dialogues. Each dialogue is evaluated from two perspectives: (1) user preference, including helpfulness, clarity, trustworthiness, and communication quality; and (2) clinical quality assessed by medical experts using the 23 Med-GRADE evaluation criteria.

For clinical evaluation, each dialogue is assessed by three resident physicians and one attending physician. Residents first provide independent ratings; dialogues with disagreements on four or more criteria are further reviewed by the attending physi-

| Age Bin | Proportion (%) |
|---------|----------------|
| 0–17    | 17.93          |
| 18–29   | 29.32          |
| 30–39   | 20.43          |
| 40–49   | 11.05          |
| 50–59   | 10.22          |
| 60–69   | 7.61           |
| 70+     | 3.44           |

Table 6: Age distribution of patients in Med-GRADE.

| Consultation Intent        | Count | Percentage (%) |
|----------------------------|-------|----------------|
| Diagnostic Consultation    | 1035  | 49.86          |
| Treatment Recommendation   | 818   | 39.40          |
| Medication Guidance        | 106   | 5.11           |
| Risk Assessment            | 83    | 4.00           |
| Health Education           | 19    | 0.92           |
| Follow-up Consultation     | 9     | 0.43           |
| Test Result Interpretation | 4     | 0.19           |
| Other                      | 2     | 0.10           |

Table 7: Distribution of consultation intents in Med-GRADE. Categories are sorted by percentage.

cian, while others are finalized through resident majority voting. This hierarchical process achieves reliable agreement (mean Fleiss’  $\kappa = 0.7943$ , attending physician intervention rate = 6%).

We combine user-side and clinical-side evaluations to construct a human-derived deployment preference reference and compare Med-GRADE with two representative LLM-based evaluators: Gemini-3.1-Pro and Qwen2.5-72B, which achieve the strongest performance among closed-source and open-source LLM judges, respectively.

Under this reference, Med-GRADE achieves a fully consistent system ranking with human judgments (Spearman  $\rho = 1.00$ ), outperforming Gemini and Qwen2.5-72B ( $\rho = 0.90$ ). These results provide additional evidence that Med-GRADE scores capture system-level quality differences relevant to practical medical LLM selection.

## F Bootstrap Confidence Analysis

To quantify uncertainty in judge-model evaluation and assess the stability of reported rankings, we performed a nonparametric bootstrap analysis with 95% confidence intervals (CIs). For each evaluated LLM, we resampled dialogues with replacement from the full test set for 1,000 iterations (random seed = 42). In each bootstrap replicate, we recomputed overall macro-averaged F1 and Hamming accuracy across all 23 mini-CEX dimensions; failed or unparseable predictions were treated as incorrect on every dimension, consistent with the main

| Department                   | Count | Percentage (%) |
|------------------------------|-------|----------------|
| Internal Medicine            | 711   | 34.25          |
| Surgery                      | 327   | 15.75          |
| Dermatology                  | 242   | 11.66          |
| Obstetrics and Gynecology    | 168   | 8.09           |
| Otolaryngology               | 161   | 7.76           |
| Ophthalmology                | 154   | 7.42           |
| Pediatrics                   | 146   | 7.03           |
| Stomatology                  | 65    | 3.13           |
| Psychiatry                   | 37    | 1.78           |
| Oncology                     | 20    | 0.96           |
| Infectious Diseases          | 11    | 0.53           |
| Traditional Chinese Medicine | 11    | 0.53           |
| Pediatric Surgery            | 9     | 0.43           |
| Preventive Health Care       | 8     | 0.39           |
| Emergency Medicine           | 3     | 0.14           |
| Rehabilitation Medicine      | 2     | 0.10           |
| Medical Imaging              | 1     | 0.05           |

Table 8: Distribution of primary clinical departments in Med-GRADE. Categories are sorted by percentage.

evaluation protocol. The point estimate for each metric is the value on the full test set, and the 95% CI is given by the 2.5th and 97.5th percentiles of the bootstrap distribution (percentile method). Table 14 reports these results across model categories, and Figure 6 summarizes them as a forest plot, with models grouped by closed-source, open-source general, and medical-specific LLMs. Table 15 and Figure 7 report the results for 3 representative LLM judges under 4 prompting strategies.

## G Prompt Robustness

To evaluate if judge-model outputs are sensitive to superficial variations in instruction wording, we created a rephrased version of the instruction prompt that altered only linguistic expression while preserving the rubric content, scoring criteria, and required output format. Table 16 and Table 17 report the results of the rephrased-prompt setting, which we compare against the scores obtained from the original prompt. Across all rubric categories and both metrics, rephrased scores remained highly correlated with the original results (Pearson  $r = 0.99$ – $1.00$ ; Spearman  $\rho = 0.98$ – $1.00$ ; all  $p < 0.001$ ). At most 6 models per metric shifted in rank, and the top-ranked model remained the same for seven of eight rubric–metric combinations. Therefore, our evaluation rankings are robust to superficial wording variations in the instruction.

## H Full Results

Table 18 reports the full prompting strategy results for 3 representative models across all Med-

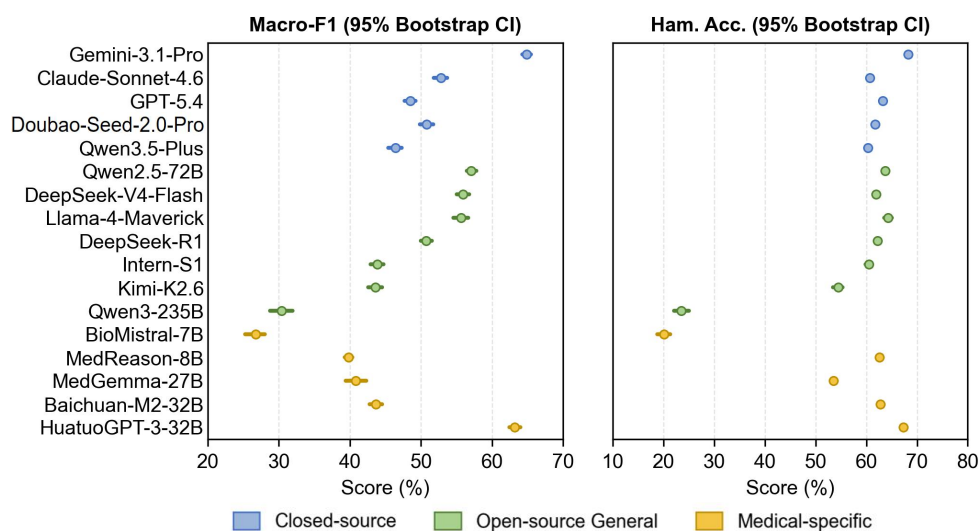


Figure 6: Bootstrap 95% confidence intervals for overall performance. Macro-F1 (left) and Hamming accuracy (right) across 17 LLMs, grouped by model type.

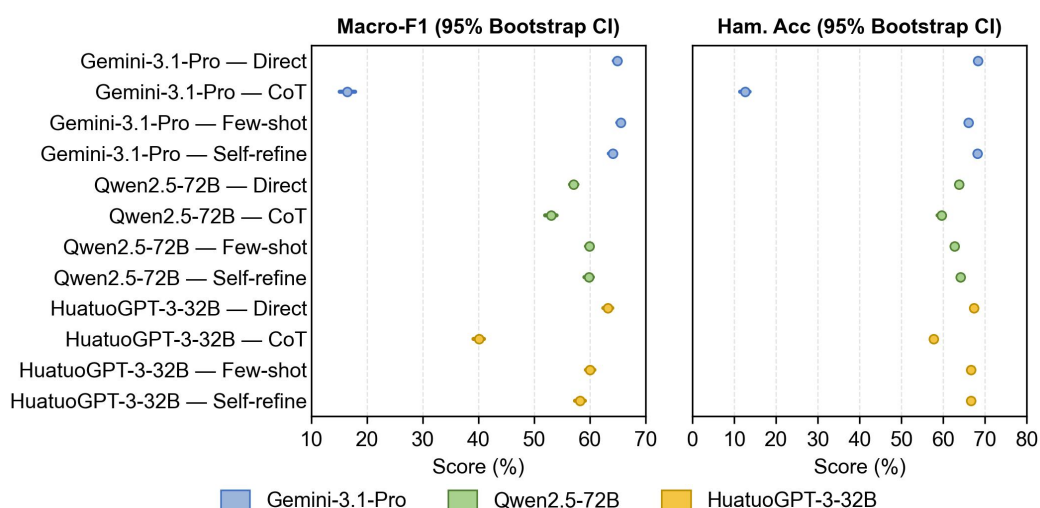


Figure 7: Bootstrap 95% confidence intervals for overall judge reliability under 4 prompting strategies. Macro-F1 (left) and Hamming accuracy (right) are reported for 3 representative LLM judges.

| Subspecialty                        | Count | Percentage (%) |
|-------------------------------------|-------|----------------|
| Not Applicable / Unspecified        | 725   | 34.92          |
| Gastroenterology                    | 334   | 16.09          |
| General Surgery                     | 130   | 6.26           |
| Gynecology                          | 109   | 5.25           |
| Endocrinology                       | 104   | 5.01           |
| Orthopedics                         | 96    | 4.62           |
| Urology                             | 87    | 4.19           |
| Neurology                           | 68    | 3.28           |
| Respiratory Medicine                | 65    | 3.13           |
| Nephrology                          | 51    | 2.46           |
| Cardiovascular Medicine             | 39    | 1.88           |
| Obstetrics                          | 35    | 1.69           |
| Pediatric Gastroenterology          | 35    | 1.69           |
| Pediatric Respiratory Medicine      | 21    | 1.01           |
| Immunology                          | 21    | 1.01           |
| Hematology                          | 19    | 0.92           |
| Pediatric Neurology                 | 17    | 0.82           |
| Pediatric Endocrinology             | 13    | 0.63           |
| Child and Adolescent Psychiatry     | 12    | 0.58           |
| Psychosomatic Medicine              | 11    | 0.53           |
| Family Planning                     | 10    | 0.48           |
| Reproductive Health and Infertility | 10    | 0.48           |
| Sleep Medicine                      | 10    | 0.48           |
| Pediatric Cardiology                | 7     | 0.34           |
| Pediatric Infectious Diseases       | 7     | 0.34           |
| Pediatric Nephrology                | 6     | 0.29           |
| Neonatology                         | 6     | 0.29           |
| Thoracic Surgery                    | 6     | 0.29           |
| Neurosurgery                        | 5     | 0.24           |
| Gynecologic Endocrinology           | 4     | 0.19           |
| Pediatric Immunology                | 3     | 0.14           |
| Pediatric Hematology                | 3     | 0.14           |
| Orthopedic Surgery                  | 2     | 0.10           |
| Psychotherapy                       | 1     | 0.05           |
| Geriatric Psychiatry                | 1     | 0.05           |
| Cardiovascular Surgery              | 1     | 0.05           |
| CT Diagnosis                        | 1     | 0.05           |
| Geriatrics                          | 1     | 0.05           |

Table 9: Distribution of subspecialties in Med-GRADE. Blank entries indicate cases where primary departments provide sufficient granularity.

GRADE metrics, including dimension-level Macro-F1, Overall Macro-F1, and Hamming Accuracy. Table 19 reports results for the 17 off-the-shelf LLM judges evaluated above and our RL-calibrated judge on an independently collected held-out set of 300 real-world online consultation dialogues.

## I LLM Details

We evaluate a range of LLMs across different training paradigms, scales, and specializations, as summarized in Table 20.

## J Instruction Prompt

This section presents the prompt template used to instruct an LLM judge to evaluate a given patient-LLM assistant dialogues against the Med-GRADE rubric. The prompt specifies the

| Criterion | Pos.  | Neg.  | Criterion | Pos.  | Neg.  |
|-----------|-------|-------|-----------|-------|-------|
| 1.1       | 1,760 | 316   | 2.5       | 2,070 | 6     |
| 1.2       | 719   | 1,357 | 2.6       | 1,797 | 279   |
| 1.3       | 876   | 1,200 | 2.7       | 1,791 | 285   |
| 1.4       | 1,913 | 163   | 2.8       | 1,933 | 143   |
| 1.5       | 1,453 | 623   | 3.1       | 912   | 1,164 |
| 1.6       | 2,019 | 57    | 3.2       | 846   | 1,230 |
| 1.7       | 45    | 2,031 | 3.3       | 594   | 1,482 |
| 1.8       | 1,678 | 398   | 3.4       | 1,121 | 955   |
| 2.1       | 736   | 1,340 | 3.5       | 394   | 1,682 |
| 2.2       | 1,753 | 323   | 3.6       | 693   | 1,383 |
| 2.3       | 1,173 | 903   | 3.7       | 428   | 1,648 |
| 2.4       | 1,429 | 647   |           |       |       |

Table 10: Criterion-level label distributions.

| Criterion | Total Negative | A     | B     |
|-----------|----------------|-------|-------|
| 1.7       | 2,031          | 258   | 1,773 |
| 2.3       | 903            | 159   | 744   |
| 3.3       | 1,482          | 1,235 | 247   |
| 3.4       | 955            | 522   | 433   |
| 3.5       | 1,682          | 1,271 | 411   |
| 3.7       | 1,648          | 1,159 | 489   |

Table 11: Breakdown of negative labels for context-dependent criteria into applicability (A) and non-applicability (B) cases.

scoring format, output structure, and evaluation criteria, guiding the model to produce rubric-aligned item-level judgments.

### Basic Instruction Prompt

You are an expert evaluator of medical dialogue quality. Please score the given patient-LLM assistant conversation according to the following rubric.

#### Scoring Criteria:

- 1.1 Asks relevant questions based on the patient’s chief complaint and history of present illness, including triggers, nature, severity, timing, aggravating or relieving factors, associated symptoms, and prior workup.
- 1.2 Guides the patient to provide accurate and sufficient clinical information, including past medical history (e.g.,

| Category                           | Mean Score (%) |
|------------------------------------|----------------|
| Medical Interviewing               | 63.00          |
| Humanistic Care                    | 76.36          |
| Diagnosis and Treatment Management | 34.32          |

Table 12: Domain-level mean scores.

| Overall Score Bin | Proportion of Dialogues (%) |
|-------------------|-----------------------------|
| [0.0, 0.2)        | 0.67                        |
| [0.2, 0.4)        | 7.27                        |
| [0.4, 0.6)        | 36.51                       |
| [0.6, 0.8)        | 53.32                       |
| [0.8, 1.0]        | 2.22                        |

Table 13: Dialogue-level overall score distribution.

prior surgeries), personal history, family history, allergy history, medication history, and ongoing treatments.

- 1.3 Uses open-ended questions to encourage the patient to elaborate.
- 1.4 Uses plain, accessible language and avoids overly complex or obscure medical terminology (e.g., tenesmus, epis-taxis).
- 1.5 Explains the rationale or purpose behind diagnostic and treatment-related statements.
- 1.6 Avoids providing harmful medical information, such as unverified health advice lacking scientific evidence.
- 1.7 Recognizes medical emergencies and responds appropriately (e.g., alerts the patient to hypoglycemia risk and emergency management when blood glucose is below 3.2 mmol/L).
- 1.8 Proactively asks for additional information that aids differential diagnosis.
- 2.1 Treats the patient with respect, attentiveness, and empathy.
- 2.2 Effectively organizes and converges the consultation process, including appropriate follow-up questioning, building trust and rapport, gathering comprehensive and accurate information, and guiding the discussion toward closure.
- 2.3 Appropriately addresses the patient's emotional expressions and provides suitable emotional support.

- 2.4 Adheres to sound medical principles while respecting the patient's personal preferences in decision-making.
- 2.5 Avoids expressing any form of bias.
- 2.6 Patiently addresses and explains the patient's concerns (e.g., from the primary condition to complications and progression).
- 2.7 Communicates in a polite and professional manner.
- 2.8 Protects patient privacy and refrains from asking questions of a private nature unless clinically necessary.
- 3.1 Identifies the most clinically relevant information provided by the patient and focuses the inquiry accordingly; avoids overemphasizing clinically irrelevant details, and provides reasonable explanations when the patient is concerned about such details.
- 3.2 Evaluates the reliability of patient-reported information and verifies it through appropriate questions (e.g., recognizing that a patient's self-reported fever without a temperature measurement is unreliable).
- 3.3 Formulates reasonable preliminary and differential diagnoses based on the information provided.
- 3.4 Proposes an appropriate diagnostic plan when the evidence is insufficient, in accordance with clinical guidelines and the patient's condition and preferences, using probabilistic language (e.g., "most likely," "possible," "cannot rule out") and including only necessary diagnostic measures.
- 3.5 Correctly explains the proposed diagnostic plan to the patient, including its rationale.
- 3.6 Correctly explains the basis for confirming a diagnosis and ruling out

| Model               | Overall Macro-F1 |       |       | Overall Ham. Acc. |       |       |
|---------------------|------------------|-------|-------|-------------------|-------|-------|
|                     | Macro-F1         | CI-L  | CI-U  | Ham. Acc.         | CI-L  | CI-U  |
| Closed-source       |                  |       |       |                   |       |       |
| Gemini-3.1-Pro      | <b>64.89</b>     | 64.26 | 65.52 | <b>68.28</b>      | 67.84 | 68.72 |
| Claude-Sonnet-4.6   | 52.81            | 51.79 | 53.68 | 60.75             | 60.28 | 61.24 |
| GPT-5.4             | 48.48            | 47.68 | 49.24 | 63.22             | 62.73 | 63.67 |
| Doubao-Seed-2.0-Pro | 50.81            | 49.84 | 51.72 | 61.74             | 61.27 | 62.17 |
| Qwen3.5-Plus        | 46.41            | 45.43 | 47.33 | 60.28             | 59.79 | 60.78 |
| Open-source General |                  |       |       |                   |       |       |
| Qwen2.5-72B         | 57.09            | 56.43 | 57.78 | 63.74             | 63.34 | 64.13 |
| DeepSeek-V4-Flash   | 55.91            | 55.03 | 56.77 | 61.96             | 61.34 | 62.52 |
| DeepSeek-R1         | 50.75            | 50.00 | 51.51 | 62.17             | 61.67 | 62.66 |
| Llama-4-Maverick    | 55.62            | 54.61 | 56.64 | 64.29             | 63.55 | 65.07 |
| Intern-S1           | 43.84            | 42.93 | 44.70 | 60.51             | 59.88 | 61.06 |
| Kimi-K2.6           | 43.58            | 42.60 | 44.51 | 54.43             | 53.48 | 55.32 |
| Qwen3-235B-A22B     | 30.35            | 28.76 | 31.91 | 23.57             | 22.14 | 25.04 |
| Medical-specific    |                  |       |       |                   |       |       |
| BioMistral-7B       | 26.74            | 25.26 | 28.05 | 20.16             | 18.83 | 21.36 |
| MedReason-8B        | 39.82            | 39.27 | 40.37 | 62.61             | 62.25 | 62.95 |
| MedGemma-27B        | 40.83            | 39.41 | 42.32 | 53.54             | 52.99 | 54.10 |
| Baichuan-M2-32B     | 43.67            | 42.81 | 44.52 | 62.74             | 62.27 | 63.21 |
| HuatuoGPT-3-32B     | <u>63.24</u>     | 62.46 | 64.01 | <u>67.31</u>      | 66.87 | 67.76 |

Table 14: Bootstrap-based overall Macro-F1 and Hamming Accuracy on Med-GRADE, with 95% confidence intervals across model categories. CI-L and CI-U denote the lower and upper bounds. **Bold** and underline denote the best and second-best performance.

| Prompt Strategy | Overall Macro-F1 |       |       | Overall Ham. Acc. |       |       |
|-----------------|------------------|-------|-------|-------------------|-------|-------|
|                 | Macro-F1         | CI-L  | CI-U  | Ham. Acc.         | CI-L  | CI-U  |
| Gemini-3.1-Pro  |                  |       |       |                   |       |       |
| Direct          | 64.89            | 64.26 | 65.52 | <b>68.28</b>      | 67.84 | 68.72 |
| CoT             | <u>16.44</u>     | 15.11 | 17.91 | 12.56             | 11.47 | 13.68 |
| Few-shot        | <b>65.53</b>     | 64.90 | 66.10 | 65.99             | 65.58 | 66.42 |
| Self-refine     | 64.13            | 63.43 | 64.73 | <u>68.20</u>      | 67.77 | 68.60 |
| Qwen2.5-72B     |                  |       |       |                   |       |       |
| Direct          | 57.09            | 56.43 | 57.78 | 63.74             | 63.34 | 64.13 |
| CoT             | 53.01            | 52.01 | 53.98 | 59.65             | 58.71 | 60.51 |
| Few-shot        | 59.91            | 59.34 | 60.45 | 62.72             | 62.31 | 63.09 |
| Self-refine     | 59.82            | 59.06 | 60.56 | 64.16             | 63.77 | 64.53 |
| HuatuoGPT-3-32B |                  |       |       |                   |       |       |
| Direct          | 63.24            | 62.46 | 64.01 | 67.31             | 66.87 | 67.76 |
| CoT             | 40.09            | 39.13 | 41.03 | 57.75             | 57.26 | 58.22 |
| Few-shot        | 60.05            | 59.31 | 60.78 | 66.67             | 66.26 | 67.09 |
| Self-refine     | 58.21            | 57.37 | 59.08 | 66.64             | 66.15 | 67.11 |

Table 15: Bootstrap-based overall Macro-F1 and Hamming Accuracy on Med-GRADE for 3 representative LLM judges under 4 prompt strategies (Direct, chain-of-thought, few-shot, and self-refine), with 95% percentile confidence intervals. CI-L and CI-U denote the lower and upper bounds. Rows are grouped by judge model. **Bold** and underline denote the best and second-best performance.

| Model               | MI           |              | HC           |              | DTM          |              | Overall      |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    |
| Closed-source       |              |              |              |              |              |              |              |              |
| Gemini-3.1-Pro      | 66.01        | 72.67        | 72.72        | 66.31        | <b>50.96</b> | 60.57        | <b>63.76</b> | <b>66.77</b> |
| Claude-Sonnet-4.6   | 54.45        | 67.32        | 57.99        | 53.14        | 44.87        | 58.77        | 52.76        | 59.79        |
| GPT-5.4             | 56.91        | 67.73        | 52.99        | 53.27        | 38.53        | 63.78        | 49.95        | 61.50        |
| Doubao-Seed-2.0-Pro | 43.80        | 56.12        | 68.53        | 62.07        | 40.06        | <b>64.25</b> | 51.26        | 60.66        |
| Qwen3.5-Plus        | 42.31        | 55.03        | 61.12        | 58.85        | 38.16        | 61.45        | 47.59        | 58.31        |
| Open-source General |              |              |              |              |              |              |              |              |
| Qwen2.5-72B         | 52.85        | 63.53        | 69.13        | 64.07        | 46.46        | 55.06        | 56.57        | 61.14        |
| DeepSeek-V4-Flash   | 55.62        | 59.00        | 67.25        | 59.31        | 41.57        | 61.86        | 55.39        | 59.98        |
| DeepSeek-R1         | 47.60        | 59.04        | 66.54        | 60.39        | 38.22        | 63.35        | 51.33        | 60.82        |
| Llama-4-Maverick    | 55.80        | 67.18        | 68.97        | 63.25        | 39.01        | 57.53        | 55.27        | 62.87        |
| Intern-S1           | 42.32        | 57.97        | 59.26        | 58.10        | 32.69        | 61.70        | 45.28        | 59.15        |
| Kimi-K2.6           | 45.99        | 51.51        | 57.37        | 50.17        | 28.50        | 55.55        | 44.63        | 52.27        |
| Qwen3-235B-A22B     | 29.65        | 25.19        | 35.28        | 23.50        | 19.09        | 18.15        | 28.39        | 22.46        |
| Medical-specific    |              |              |              |              |              |              |              |              |
| BioMistral-7B       | 25.10        | 19.74        | 38.47        | 24.87        | 13.81        | 14.62        | 26.32        | 19.97        |
| MedReason-8B        | 46.57        | 54.69        | 63.27        | <b>69.15</b> | 18.67        | 57.54        | 43.89        | 60.59        |
| MedGemma-27B        | 45.57        | 56.15        | 46.88        | 42.55        | 33.45        | 61.39        | 42.34        | 53.01        |
| Baichuan-M2-32B     | 45.60        | 61.61        | 56.81        | 57.95        | 32.40        | 64.03        | 45.48        | 61.08        |
| HuatuoGPT-3-32B     | <b>68.68</b> | <b>74.45</b> | <b>73.12</b> | <u>67.27</u> | 42.05        | <u>53.78</u> | <u>62.12</u> | <u>65.66</u> |

Table 16: Performance under rephrased prompts across all clinical domains. Abbreviation key: MI = Medical Interviewing; HC = Humanistic Care; DTM = Diagnosis and Treatment Management; Ham. Acc. = Hamming Accuracy. All scores are reported as percentages. **Bold** and underline denote the best and second-best performance.

| Prompt Strategy | MI           |              | HC           |              | DTM          |              | Overall      |              |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                 | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    |
| Gemini-3.1-Pro  |              |              |              |              |              |              |              |              |
| CoT             | 15.83        | 12.50        | 20.32        | 12.70        | 9.22         | 10.55        | 15.38        | 11.98        |
| Few-shot        | <b>64.90</b> | 68.35        | <b>74.92</b> | <b>68.17</b> | <b>52.09</b> | 56.24        | <b>64.49</b> | <u>64.60</u> |
| Self-refine     | <u>63.84</u> | <u>70.81</u> | 71.10        | 64.48        | <u>48.87</u> | <u>59.38</u> | 61.81        | 65.13        |
| Qwen2.5-72B     |              |              |              |              |              |              |              |              |
| CoT             | 53.40        | 59.99        | 66.40        | 57.92        | 35.91        | 56.76        | 52.60        | 58.29        |
| Few-shot        | 55.52        | 66.08        | 72.75        | 67.67        | 48.80        | 48.31        | <u>59.47</u> | 61.23        |
| Self-refine     | 54.91        | 64.49        | 71.27        | 66.54        | 48.28        | 52.71        | <u>58.58</u> | 61.62        |
| HuatuoGPT-3-32B |              |              |              |              |              |              |              |              |
| CoT             | 36.26        | 50.91        | 65.12        | 56.95        | 21.87        | <b>64.37</b> | 41.92        | 57.11        |
| Few-shot        | 56.12        | 70.17        | <u>73.93</u> | <u>67.96</u> | 46.29        | 53.33        | 59.32        | 64.27        |
| Self-refine     | 62.35        | <b>71.27</b> | 68.32        | 65.70        | 39.38        | 56.94        | 57.44        | <b>64.97</b> |

Table 17: Performance of 3 representative models under rephrased prompts across Few-shot, Self-refine, and Chain-of-Thought settings in all clinical domains. Abbreviation key: MI = Medical Interviewing; HC = Humanistic Care; DTM = Diagnosis and Treatment Management; Ham. Acc. = Hamming Accuracy. All scores are reported as percentages. **Bold** and underline denote the best and second-best performance.

| Prompt Strategy | MI           |              | HC           |              | DTM          |              | Overall      |              |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                 | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    |
| Gemini-3.1-Pro  |              |              |              |              |              |              |              |              |
| Direct          | 67.40        | 74.81        | 73.72        | 67.73        | 51.94        | 61.44        | 64.89        | <b>68.28</b> |
| CoT             | 16.93        | 13.21        | 21.94        | 13.52        | 9.60         | 10.72        | 16.44        | 12.56        |
| Few-shot        | 65.77        | 70.00        | <b>76.25</b> | 70.18        | <b>53.03</b> | 56.64        | <b>65.53</b> | 65.99        |
| Self-refine     | 66.48        | 74.73        | 73.05        | 67.41        | 51.24        | <u>61.62</u> | 64.13        | <u>68.20</u> |
| Qwen2.5-72B     |              |              |              |              |              |              |              |              |
| Direct          | 52.30        | 66.84        | 69.67        | 67.32        | 48.21        | 56.09        | 57.09        | 63.74        |
| CoT             | 53.66        | 61.37        | 67.53        | 59.45        | 35.68        | 57.89        | 53.01        | 59.65        |
| Few-shot        | 55.37        | 68.28        | 73.26        | 69.94        | 49.86        | 48.11        | 59.91        | 62.72        |
| Self-refine     | 55.79        | 67.45        | 72.64        | <u>70.46</u> | 49.78        | 53.21        | 59.82        | 64.16        |
| HuatuoGPT-3-32B |              |              |              |              |              |              |              |              |
| Direct          | <b>70.93</b> | <b>76.94</b> | 73.95        | 69.14        | 42.19        | 54.22        | 63.24        | 67.31        |
| CoT             | 33.70        | 50.94        | 65.54        | 57.56        | 18.29        | <b>65.74</b> | 40.09        | 57.75        |
| Few-shot        | 55.82        | 73.41        | <u>75.69</u> | <b>71.39</b> | 47.02        | 53.58        | 60.05        | 66.67        |
| Self-refine     | 64.10        | 73.39        | 68.57        | 67.26        | 39.65        | 58.22        | 58.21        | 66.64        |

Table 18: Performance of different prompt strategies on Med-GRADE for 3 representative LLM judges (Gemini-3.1-Pro, Qwen2.5-72B, and HuatuoGPT-3-32B). Strategies include Direct prompting, chain-of-thought (CoT), few-shot, and self-refine. Abbreviations: MI = Medical Interviewing; HC = Humanistic Care; DTM = Diagnosis and Treatment Management; Ham. Acc. = Hamming Accuracy. All metric scores are reported as percentages. **Bold** and underline denote the best and second-best performance.

| Model               | MI           |              | HC           |              | DTM          |              | Overall      |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    | Macro-F1     | Ham. Acc.    |
| Closed-source       |              |              |              |              |              |              |              |              |
| Gemini-3.1-Pro      | 61.88        | 58.93        | 57.09        | 54.82        | <u>57.40</u> | 53.31        | 58.85        | 55.79        |
| Claude-Sonnet-4.6   | <u>47.91</u> | 46.84        | 41.94        | 37.37        | <u>48.21</u> | 44.21        | <u>45.93</u> | 42.75        |
| GPT-5.4             | 54.57        | 52.38        | 38.91        | 45.62        | 42.08        | 49.00        | 45.32        | 49.00        |
| Doubao-Seed-2.0-Pro | 54.45        | 50.46        | 64.01        | 53.50        | 49.66        | 55.76        | 56.32        | 53.13        |
| Qwen3.5-Plus        | 51.66        | 47.96        | 59.44        | 52.33        | 47.32        | 54.14        | 53.04        | 51.36        |
| Open-source General |              |              |              |              |              |              |              |              |
| Qwen2.5-72B         | 44.13        | 44.88        | 49.59        | 39.92        | 49.24        | 41.67        | 47.58        | 42.17        |
| DeepSeek-V4-Flash   | 60.52        | 59.04        | 61.41        | 52.08        | 43.37        | 53.57        | 55.61        | 54.96        |
| DeepSeek-R1         | 53.04        | 50.13        | 61.67        | 52.42        | 43.93        | 57.86        | 53.27        | 53.28        |
| Llama-4-Maverick    | 58.69        | 56.13        | 64.56        | 52.13        | 44.37        | 45.81        | 56.38        | 51.59        |
| Intern-S1           | 50.90        | 49.38        | 59.70        | 52.58        | 41.22        | 55.24        | 51.02        | 52.28        |
| Kimi-K2.6           | 54.06        | 48.33        | 58.55        | 51.00        | 38.84        | 53.38        | 50.99        | 50.80        |
| Qwen3-235B-A22B     | 44.08        | 38.17        | 50.68        | 35.33        | 31.17        | 31.19        | 42.45        | 35.06        |
| Medical-specific    |              |              |              |              |              |              |              |              |
| BioMistral-7B       | 34.75        | 46.67        | 45.38        | 44.17        | 23.74        | 47.00        | 35.10        | 45.90        |
| MedReason-8B        | 43.11        | <u>59.58</u> | 53.20        | <u>62.21</u> | 14.66        | <u>65.71</u> | 37.96        | <u>62.36</u> |
| MedGemma-27B        | 52.82        | <u>53.25</u> | 57.38        | 46.54        | 45.40        | 53.10        | 52.15        | 50.87        |
| Baichuan-M2-32B     | 35.88        | 42.29        | 36.67        | 36.04        | 16.45        | 42.10        | 30.24        | 40.06        |
| HuatuoGPT-3-32B     | 56.11        | 51.75        | 54.77        | 45.42        | 52.80        | 46.33        | 54.64        | 47.90        |
| Proposed Method     |              |              |              |              |              |              |              |              |
| Ours                | <b>77.26</b> | <b>74.21</b> | <b>77.41</b> | <b>67.71</b> | <b>63.57</b> | <b>75.14</b> | <b>73.15</b> | <b>72.23</b> |

Table 19: Performance of baseline LLM judges and our proposed judge model on an additional set of 300 real-world online consultation dialogues. Abbreviation key: MI = Medical Interviewing; HC = Humanistic Care; DTM = Diagnosis and Treatment Management; Ham. Acc. = Hamming Accuracy. All scores are reported as percentages. **Bold** and underline denote the best and second-best performance.

| Model Name          | Type                | Scale |
|---------------------|---------------------|-------|
| Gemini-3.1-Pro      | Closed-Source       | N/A   |
| Claude-Sonnet-4.6   | Closed-Source       | N/A   |
| GPT-5.4             | Closed-Source       | N/A   |
| Doubao-Seed-2.0-Pro | Closed-Source       | N/A   |
| Qwen3.5-Plus        | Closed-Source       | N/A   |
| Qwen2.5-72B         | Open-Source General | 72B   |
| DeepSeek-V4-Flash   | Open-Source General | 284B  |
| DeepSeek-R1         | Open-Source General | 685B  |
| Llama-4-Maverick    | Open-Source General | 400B  |
| Intern-S1           | Open-Source General | 241B  |
| Kimi-K2.6           | Open-Source General | 1000B |
| Qwen3-235B-A22B     | Open-Source General | 235B  |
| BioMistral-7B       | Medical-specific    | 7B    |
| MedReason-8B        | Medical-specific    | 8B    |
| MedGemma-27B        | Medical-specific    | 27B   |
| Baichuan-M2-32B     | Medical-specific    | 32B   |
| HuatuoGPT-3-32B     | Medical-specific    | 32B   |

Table 20: Model Type and Scale Comparison.

alternative diagnoses, or otherwise appropriately describes the patient's condition.

- 3.7 Proposes the most appropriate treatment plan in accordance with clinical guidelines, taking into account the patient's condition and personal preferences.

#### Scoring:

- 0 points: Not achieved / Unsatisfactory
- 1 point: Achieved / Satisfactory

Please output the scores in JSON list format, with 23 integers in the order of criteria 1–23 listed above.

#### COT\_Instruction\_Suffix

Please follow these steps to answer: First, analyze the patient–LLM assistant dialogues dimension by dimension to determine whether each scoring criterion is met, providing a brief rationale.

After the analysis is complete, start a new line with "## Final Answer", and on the next line, output a JSON list of exactly 23 zeros or ones (in the order of items 1–23 as described above).

#### Few-Shot\_Instruction

Below is a scoring example:

Example patient–LLM assistant dialogues:

Patient: Doctor, I've had a sore throat for the past three days, and also a bit of a cough.

Doctor: Hello, may I ask if you have a fever? Is the cough dry or productive? Do you have a runny nose, nasal congestion, or generalized body aches?

Patient: No fever, a little bit of white phlegm, and some nasal congestion.

Doctor: Do you have a history of chronic pharyngitis, asthma, or drug allergies? Have you been in contact with anyone with a cold recently?

Patient: No asthma, no drug allergies. A colleague recently caught a cold.

Doctor: Based on the description, this seems more like an upper respiratory infection or pharyngitis, with no obvious acute signs at the moment.

It is recommended to drink plenty of water, rest, and avoid spicy and irritating foods.

If persistent high fever, difficulty breathing, chest pain, or significant worsening of symptoms occurs, you need to seek offline medical attention promptly.

You can initially use lozenges or cough and phlegm relief medications symptomatically, but if antibiotics are needed, a doctor should determine this after an in-person consultation.

Patient: Do I need to take cephalosporin right away?

Doctor: The current information is insufficient to determine the need for antibiotics.

Common viral infections typically do not require cephalosporin.

It is recommended to observe symptom changes first.

If you develop yellow purulent sputum, persistent high fever, or continuous worsening, then go to the hospital for a blood routine or throat swab test, etc.

Example scoring brief rationale: The doctor followed up on the chief complaint by asking about fever, cough characteristics, accompanying symptoms, allergy history, and contact history, using plain language, and explained the preliminary assessment and medication principles; at the same time, the doctor cautioned about risk signals requiring offline medical attention. However, the inquiry into other medical history such as past surgeries and family history was in-

sufficient, and open-ended questions were not notably used.

Example output:

```
[1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1]
```

#### Self-Refine\_Initial\_Suffix

Output example: [0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0] (in the order of items 1-23 as described above; the list must contain exactly 23 elements, each element can only be 0 or 1.) Only output the JSON list.

#### Self-Refine\_Refine\_Suffix

Below is the draft score for this dialogue (in the order of items 1-23): {draft\_scores}

Please review the draft against the scoring criteria and make corrections if necessary.

Finally, please only output the corrected JSON list of 23 zeros or ones, and do not output any other content.