

MedEQ: A Clinical Emotional Intelligence Benchmark and Rubric-Guided Reinforcement Learning for Patient-Centric Medical LLMs

Yiming Zhang^{1,2†}, Xiaoyan Yang^{2†}, Yanchao Xu^{1†}, Jiayi Deng^{2†}

Yingfan Ma², Lirong Gao¹, Yanmei Gu², Kaiya Zheng²

Bowen Song², Jinjie Gu², Junbo Zhao^{1,2*}

[†]Equal contribution. ^{*}Corresponding author.

¹Zhejiang University ²Ant Group

Correspondence: j.zhao@zju.edu.cn

Abstract

Patient-facing medical LLMs are typically evaluated for clinical knowledge and safety, but clinically correct answers can still fail patients by missing fear, mishandling uncertainty, or offering emotionally unsafe reassurance. We introduce MedEQ, a benchmark for clinical emotional intelligence in medical communication. MedEQ includes 413 clinically grounded consultation scenarios across 51 specialties and 7,336 expert-refined rubric items over 21 core dimensions. These expert-validated rubrics operationalize emotional cue recognition, empathic adequacy, ethical emotional regulation, and affective safety, enabling diagnostic evaluation rather than a single holistic empathy score. Using an LLM-as-judge rubric pipeline, we find that strong contemporary models often handle surface compassion, safety, and response structure, yet remain weak on deeper emotional inference, personalization, mixed emotions, cultural context, and real-time responses to illness-related distress. We further conduct a proof-of-concept rubric-guided reinforcement learning study, showing that MedEQ-derived rubric signals can transfer to an independent emotional intelligence benchmark: rubric-guided training of Qwen3-32B improves EQ-Bench performance from 42.42 to 81.49 while preserving or slightly improving standard medical benchmark scores. MedEQ¹ therefore makes clinical emotional intelligence measurable and decomposable, and provides evidence that expert-refined rubrics can support the evaluation and improvement of patient-facing medical AI.

1 Introduction

Empathy is not merely a conversational style in medical applications. It shapes whether patients disclose symptoms, understand risk, trust advice,

and follow care guidance, especially when they are anxious, ashamed, in pain, or facing uncertain outcomes (Riess, 2017; Derksen et al., 2013; Street Jr et al., 2009). For patient-facing LLMs, this means that a clinically correct answer can still fail as medical communication: it may overlook fear, respond to shame with generic reassurance, make uncertainty sound falsely resolved, or miss emotional cues that change how a patient should be supported.

Recent healthcare LLM benchmarks have moved beyond exam-style QA toward realistic multi-turn conversations scored by physicians (Arora et al., 2025; Hicks et al., 2026; Bedi et al., 2025; Tu et al., 2025). In parallel, emotional-intelligence benchmarks have expanded from emotion recognition to rubric-guided empathetic dialogue (Paech, 2023; Chen et al., 2024; Hu et al., 2025a; Yuan et al., 2025). Yet the two lines of work remain largely separate: medical benchmarks primarily assess correctness, safety, and broad communication quality without isolating emotional intelligence, while EI benchmarks lack the clinical constraints—uncertainty communication, role boundaries, reassurance without false guarantees—that make empathy in medicine distinct.

This gap matters because clinical empathy is not equivalent to warmth. A useful medical response must recognize the patient’s emotional state, explain uncertainty without over-reassurance, remain within the clinician’s role, and preserve safety when distress, misconceptions, or urgent symptoms are present. The missing piece is a benchmark that brings fine-grained, rubric-centered evaluation to the affective layer of clinical care.

To address this gap, we introduce **MedEQ**, a benchmark that makes clinical emotional intelligence the primary target of medical LLM evaluation. Here, “EQ” refers to emotional intelligence. Rather than assigning a single empathy score, MedEQ decomposes medical emotional in-

¹We have already published the code and data at <https://github.com/AQ-MedAI/MedEQBench>. The Chinese and English versions are aligned at the scenario level.

telligence into fine-grained, expert-validated rubric dimensions that score responses, diagnose failure modes, and provide structured signals for model improvement. Our contributions are:

- **MedEQ Dataset:** 413 clinically grounded consultation scenarios covering 51 specialties, capturing diverse emotional and psychological challenges without containing real patient-identifying information, with both Chinese and English versions planned for release.
- **Clinical Emotional Intelligence Rubrics:** 7,336 expert-refined rubric items across 21 core dimensions that operationalize emotional recognition, empathic adequacy, ethical regulation, and affective safety.
- **Diagnostic Evaluation:** A fine-grained evaluation framework that identifies distinct failure modes instead of collapsing medical empathy into a single holistic preference score.
- **Rubric-Guided RL Validation:** A proof-of-concept reinforcement learning study showing that MedEQ-derived rubric signals can improve performance on an independent emotional intelligence benchmark while preserving standard medical QA performance.

2 Related Work

2.1 Medical LLMs and Patient-Centric Clinical Dialogue

Medical LLM evaluation has evolved from exam-style clinical knowledge tests toward richer assessments of patient-facing behavior. Med-PaLM 2 and GPT-4 medical evaluations established strong performance on clinical reasoning, factuality, and safety-oriented expert preference judgments (Singhal et al., 2023; Nori et al., 2023), while MedDialog enabled training and evaluation on doctor-patient dialogue data across specialties (He et al., 2020). More recent benchmarks make this shift explicit: HealthBench uses realistic multi-turn health conversations with physician-written rubric criteria, HealthBench Professional evaluates real clinician chats with adjudicated physician rubrics, and MedHELM proposes a broader clinician-validated taxonomy for medical language model evaluation (Arora et al., 2025; Hicks et al., 2026; Bedi et al., 2025). Conversational diagnostic AI further evaluates interactive systems using patient-centered

clinical criteria, including communication and empathy (Tu et al., 2025; Chen et al., 2025). However, these benchmarks primarily target medical quality, diagnostic process, safety, or broad communication behavior; they do not systematically decompose clinical emotional intelligence into scenario-specific, safety-aware rubric dimensions.

2.2 Emotional Intelligence, Empathy, and Supportive Dialogue Benchmarks

Research on empathy and emotional intelligence (EI) in dialogue systems has progressed from contextually grounded empathetic response generation and large-scale emotion annotation (Rashkin et al., 2019; Demszky et al., 2020) toward broader evaluation of emotional understanding and expressed behavior. Psychometric-style benchmarks such as SECEU and EQ-Bench assess EI using standardized tests and emotion intensity prediction (Wang et al., 2023a; Paech, 2023). Recent frameworks increasingly evaluate empathy as observable response quality, including EmotionQueen, causally enriched empathetic datasets such as ECC, and large-scale synthetic empathy data construction (Chen et al., 2024; He et al., 2025; Wang et al., 2025; Liang et al., 2024). Multimodal EI benchmarks such as EmoBench-M and MME-Emotion further extend evaluation to audio-visual affective signals (Hu et al., 2025a; Zhang et al., 2025). These studies show rapid progress in general EI evaluation, but their settings typically lack the clinical constraints that make empathy in medicine distinct: uncertainty communication, role boundaries, urgent-risk framing, reassurance without false guarantees, and disease-specific emotional reasoning.

2.3 Rubric-Guided Evaluation and Alignment for Empathy

Alignment methods such as reinforcement learning from human feedback and rule-based frameworks demonstrate that structured preference signals and explicit principles can shape LLM behavior (Ouyang et al., 2022; Wang et al., 2023b; Schulman et al., 2017). Preference optimization approaches, including DPO and GRPO, further improve alignment efficiency by learning from comparative or grouped feedback (Rafailov et al., 2023; Shao et al., 2024). In evaluation, rubric-based judging has become a major direction for making LLM assessment more interpretable and scalable; recent work such as Autorubric explicitly studies rubric calibration, bias mitigation, and psychometric rela-

bility for LLM-as-judge systems (Rao and Callison-Burch, 2026). In emotional support, Kardia-R1 introduces explainable rubric-based rewards for reinforcement learning of empathetic dialogue (Yuan et al., 2025), while mental-health-oriented work argues that empathy must be evaluated together with therapeutic intent and safety (Lai et al., 2023; Hu et al., 2025b). Recent health-evaluation frameworks similarly identify rubric design as a key bottleneck for scalable and reliable assessment (Mallinar et al., 2025).

In contrast to prior work that primarily targets general-domain empathy or correctness-driven clinical benchmarks, MedEQ focuses on the affective layer of clinical dialogue. Its expert-refined rubrics enable structured evaluation of emotional recognition, empathetic adequacy, ethical emotional regulation, and affective safety, providing a principled basis for both diagnosis and optimization in patient-facing medical LLMs.

3 MedEQ Dataset

3.1 Data Source, Screening, and Statistics

MedEQ was constructed from an initial pool of approximately 5,000 real patient consultations. These consultations were neither written by the researchers nor generated by LLMs. All records were de-identified before screening by removing names, medical record numbers, contact details, and other personally identifiable information.

Three researchers independently screened the de-identified consultations and resolved disagreements through discussion. Inclusion required a patient-originated medical inquiry, sufficient contextual information, and an explicit or implicit need for emotionally appropriate communication. We excluded duplicates, non-medical or purely administrative queries, records with insufficient context, queries without a meaningful emotional-support requirement, and records that could not be reliably de-identified. Screening and subsequent rubric validation yielded the final 413 scenarios across 51 specialties.

Each instance is paired with a set of fine-grained evaluation rubrics, resulting in 7,336 rubric items, with an average of 17.8 rubric criteria per dialogue. Rather than assigning a single emotional label, MedEQ decomposes emotional intelligence into atomic, assessable components across 21 core evaluation dimensions. Additional raw annotation labels are used internally for organization and quality

control, but they are not treated as core evaluation dimensions.

Figure 1 summarizes the overall dataset, rubric-construction, and evaluation pipeline.

3.2 Dataset coverage and distribution

MedEQ includes scenarios from 51 specialties, with counts varying according to the distribution of eligible consultations after screening. Figure 2 should therefore be interpreted as showing breadth of coverage rather than uniform representation. Rubric density also varies because individual dimensions are not equally applicable to every clinical context.

3.3 MedEQ vs Others

As summarized in Table 1, MedEQ differs from existing emotional intelligence and empathy-related benchmarks along several key dimensions (Paech, 2023; Liu et al., 2025; He et al., 2025; Yuan et al., 2025). Unlike recognition-based benchmarks such as EQ-Bench and EMO-Reasoning, MedEQ evaluates models in a generation setting, reflecting realistic medical conversational use cases. Compared to prior generation-focused datasets (e.g., ECC and Kardia-R1), MedEQ provides a substantially more structured evaluation framework, with 21 explicitly defined evaluation dimensions rather than coarse or holistic scoring.

In terms of data provenance, most existing benchmarks rely on manually constructed or synthetic conversations, whereas MedEQ is grounded in real clinical service needs across 51 specialties while excluding patient-identifying information. Although MedEQ contains fewer instances than some large-scale generation datasets, its dense evaluation design enables fine-grained diagnosis of model behavior, trading horizontal scale for vertical evaluation depth. We note that recent medical evaluation frameworks such as HealthBench and AMIE do incorporate communication and empathy axes; however, their primary objectives remain medical correctness, diagnostic workflow, and broad communication quality rather than systematic, fine-grained assessment of clinical emotional intelligence. MedEQ is, to our knowledge, the first benchmark in the medical domain that makes patient emotional care its central and decomposable evaluation target.

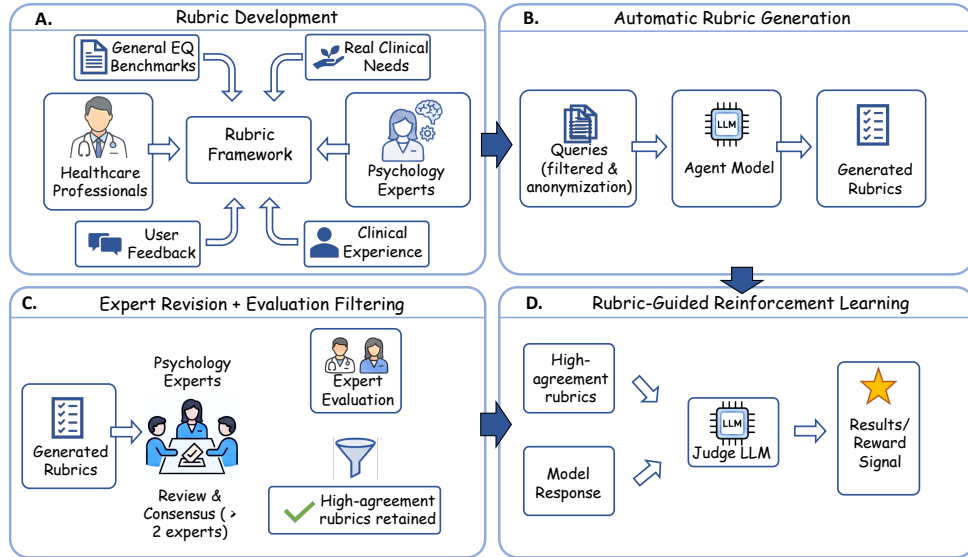


Figure 1: Overview of the MedEQ rubric construction and training pipeline. (A) **Rubric development** combines EQ benchmarks, clinical needs, expert experience, and user feedback. (B) **Automatic generation** uses an agent LLM to draft rubrics from filtered, anonymized queries. (C) **Expert verification** retains only rubrics unanimously approved after expert review. (D) **Rubric-guided reinforcement learning** converts rubric-based judge scores into reward signals for model optimization.

Table 1: Comparison between MedEQ and representative emotional intelligence and empathy-related benchmarks.

Aspect	EQ-Bench	EMO-Reasoning	ECC	Kardia-R1	MedEQ
Publication year	2023	2025	2025	2025	2025
Primary task	Recognition	Recognition	Generation	Generation	Generation
Data source	Manually constructed	Synthetic	LLM-generated	Synthetic	51 clinical specialties
Number of instances	60	~200	2,398	~22000	413
Evaluation granularity	4	~6	classification	classification	21

4 Method

4.1 Task Definition and Evaluation Formulation

We formulate emotional intelligence (EI) in clinical dialogue as a conditional language generation task grounded in realistic patient-provider interactions. Given a patient-side conversational context that may include medical information needs and implicit or explicit emotional cues (e.g., concern, distress, or uncertainty), the goal is to generate a response that is both clinically appropriate and emotionally competent. Unlike emotion recognition or classification, this task does not require predicting explicit emotional labels; instead, evaluation focuses on properties of the generated response.

Evaluation Formulation. Evaluation is framed as a rubric-based assessment to reflect the absence of a unique correct response. For each input-output pair (x, y) , responses are evaluated against a pre-defined set of rubric criteria $\{r_1, \dots, r_K\}$, each corresponding to a distinct aspect of emotional

competence in clinical communication (e.g., recognition of emotional cues, proportional empathy, or adherence to clinical norms).

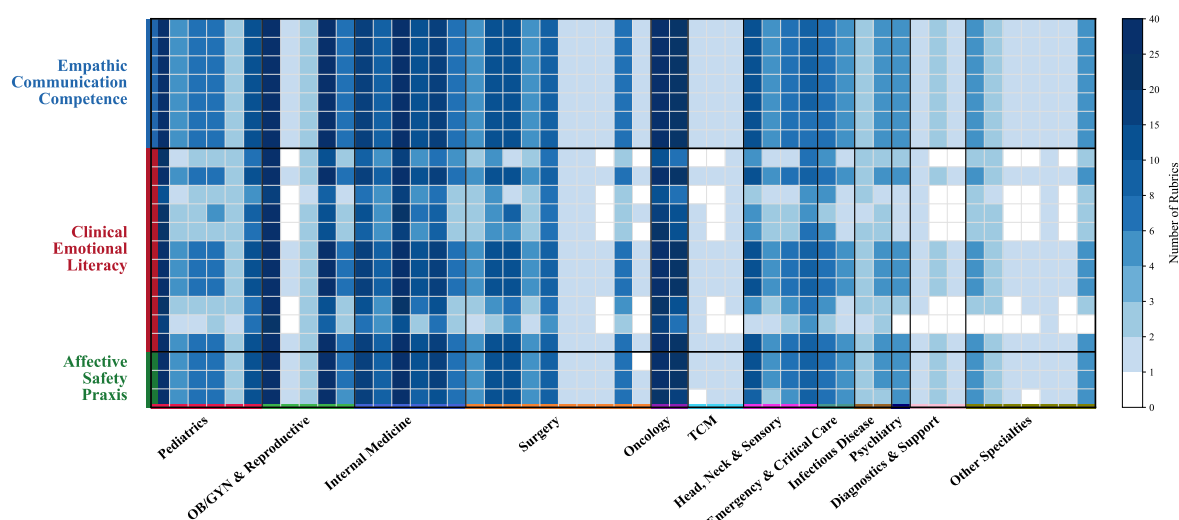
Each criterion is assessed independently, yielding a vector of rubric-level scores

$$\mathbf{s}(x, y) = (s_1, \dots, s_K).$$

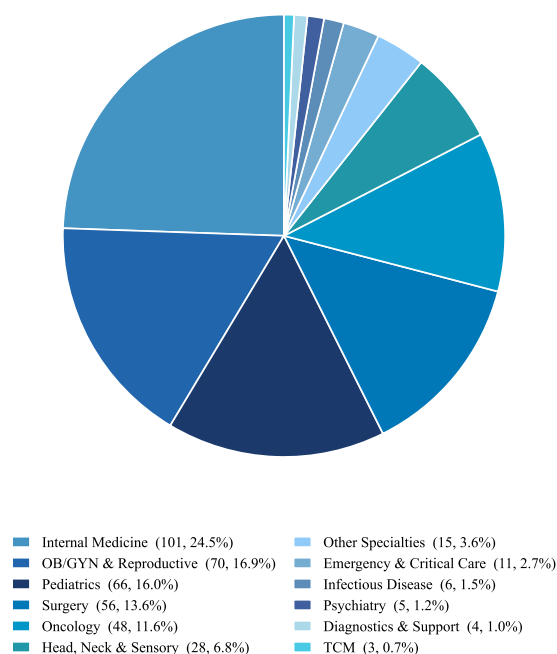
Aggregate scores can be computed across criteria or instances, but the primary signal is the structured rubric profile, which preserves fine-grained information about model behavior. This formulation enables diagnostic analysis by distinguishing different sources of failure, such as missed emotional cues or inappropriate tone, that are obscured by single-score metrics.

4.2 Rubric Design and Construction

Building on the screened scenario set described above, the rubric system is designed to operationalize emotionally competent medical communication in an interpretable and automatable manner. It



(a) Specialty–dimension coverage.



(b) Specialty distribution.

Figure 2: Specialty and rubric coverage in MedEQ. Top: distribution of scenarios across grouped clinical specialties. Bottom: coverage over 51 specialties and 21 rubric dimensions, where cell intensity denotes the number of rubric items for each specialty–dimension pair. Department group definitions are provided in Table 4.

consists of 21 evaluation dimensions grouped under three macro-competencies, with behaviorally grounded criteria used for scenario-level scoring. The taxonomy was developed through a review of emotional-intelligence and clinical-empathy frameworks, analysis of patient-facing communication needs, and iterative expert validation. The initial taxonomy contained nearly 30 candidate dimen-

sions; experts merged redundant dimensions and removed those that could not be reliably operationalized, yielding the final 21 dimensions.

For analysis, rubric dimensions are grouped into three macro-competencies: *Empathic Communication*, *Clinical Emotional Literacy*, and *Affective Safety*, which together capture complementary requirements for emotionally appropriate clinical di-

alogue. Each dimension is decomposed into minimal, observable criteria to support diagnostic evaluation and error analysis. Full rubric definitions and design rationale are provided in Appendix C.

4.3 Human-in-the-Loop Rubric Validation

As shown in Figure 1, to ensure that automatically generated rubrics were clinically meaningful and operationally precise, we employed a human-in-the-loop validation procedure centered on expert review and consensus formation.

Expert composition. Validation was conducted by a panel of 30 domain experts, comprising 12 physicians from tertiary teaching hospitals with direct patient-facing clinical experience and 18 licensed psychological counselors with training in behavioral measurement and affective assessment. This composition was intended to balance clinical plausibility and safety considerations with construct validity and annotation clarity. Experts were asked to assess whether each rubric could be applied reliably by a trained annotator using only the written definitions, without relying on implicit or contextual knowledge.

Annotation and validation process. Following initial screening, the annotation pipeline began with 500 candidate clinical queries, each paired with approximately 21 candidate rubric items generated by an LLM, yielding roughly 10,500 initial rubric entries. The validation proceeded in two stages.

In the *discussion stage*, each rubric item was reviewed by at least two experts who jointly examined construct definitions, scoring descriptions, and decision boundaries. Experts edited items to reduce ambiguity and clarify inclusion and exclusion criteria, with emphasis on conservative revisions that improved operational clarity even at the cost of increased annotation effort.

In the *verification stage*, each revised rubric item was independently assessed by three additional experts. We applied a strict unanimity rule: only rubric items accepted by all three verifiers were retained in the final set. Items rejected by any verifier were either revised and re-submitted or permanently removed.

At the query level, candidate queries whose associated rubric items exhibited a high rejection rate were discarded entirely, as such patterns typically indicated that the underlying clinical scenario was insufficiently specified to support reliable emotional intelligence evaluation.

This pipeline reduced the initial pool of 500 queries and approximately 10,500 rubric items to the final set of 413 queries and 7,336 validated rubric items, corresponding to a 30.1% rubric rejection rate under the unanimity requirement.

4.4 LLM-as-Judge Evaluation

We adopt an LLM-as-judge approach for scalable, rubric-based evaluation of model responses. For each input–output pair (x, y) , a judge language model receives the clinical scenario x , the model response y , and the corresponding rubric criteria. The judge is not given a reference answer; it independently scores whether the response satisfies each criterion.

Given a rubric with K criteria, the judge assigns a discrete score $s_k(y) \in \{0, 1, 2, 3, 4, 5\}$ to each criterion k , with an associated weight w_k . The normalized score is

$$S_{\text{norm}}(y) = \frac{\sum_{k=1}^K w_k s_k(y)}{\sum_{k=1}^K 5 w_k} \in [0, 1].$$

Here, 5 is the maximum value on the 0–5 Likert scale. We use uniform weighting throughout, with $w_k = 1$ for every rubric criterion. Outputs violating hard constraints (format, repetition, language) receive a fixed penalty score. Unless otherwise stated, all reported MedEQ scores are normalized to percentages. To validate automated scoring, 20 domain experts and 20 general-public annotators evaluated 100 stratified instances spanning high-, medium-, and low-scoring responses across multiple dimensions, yielding 3,632 rubric-level annotations. We select GPT-4.1 for its balance of expert agreement, stability, and efficiency; it also belongs to the highest-consensus judge cluster, with pairwise Kappa values of 0.736–0.784. Further details are provided in Appendices D and E.

5 Experiments

We evaluate a diverse set of state-of-the-art large language models (LLMs) on MedEQ, a benchmark designed to assess clinical emotional intelligence in generated medical consultation responses.

5.1 Experimental Setup

The evaluated models span multiple families, parameter scales, and release periods, including both proprietary and open-source systems. For models that provide explicit reasoning or “thinking” modes, we enable the provider-recommended default configuration. All models are evaluated in a zero-shot

setting without task-specific fine-tuning, external tools, or retrieval augmentation. A unified instruction template is used for all models, with identical inputs and output constraints.

We report normalized rubric score (%) on the MedEQ test set as the primary evaluation metric. For each response, the judge receives the scenario, the model response, and the corresponding rubric criteria, assigns a 0–5 score to each rubric item, and the weighted normalized score S_{norm} is computed as described in Section 4.4. Instance-level scores are averaged across the test set and reported as percentages. Invalid outputs that violate hard constraints such as language, format, or excessive repetition receive the fixed invalid-output reward before aggregation. All results are obtained using the same evaluation pipeline to minimize implementation variance. For clarity, abbreviated model names are used in tables and figures; full model identifiers and release details are provided in Appendix A.

5.2 Main Results

Figure 3 reports the performance of state-of-the-art LLMs on MedEQ.

Qwen3 achieves the best overall performance. Qwen3-235B obtains the highest normalized rubric score (55.5%), only slightly ahead of Qwen3-30B (55.3%), suggesting that scale helps but does not fully explain MedEQ performance.

Newer and reasoning-oriented models perform better. DeepSeek-V3.1 improves over DeepSeek-R1-0528 by more than 6.6 percentage points, and models with explicit reasoning mechanisms, such as Qwen3, Gemini-2.5-Pro, Baichuan-M2-32B, and DeepSeek-V3.1, outperform general-purpose models including GPT-4o and Claude-Sonnet-4.5.

GPT-family models underperform on MedEQ. This may reflect differences in training emphasis and response style: Qwen may better capture humanistic concern, whereas GPT models tend to adopt a more formal and serious tone in medical contexts. We further analyze these failure modes in the case study section.

5.3 Dimension-Level Analysis of MedEQ

Fine-Grained Dimension Analysis. We further analyze model performance across 21 core MedEQ evaluation dimensions. A clear contrast emerges between dimensions related to surface-level com-

munication quality and those requiring deep emotional inference and personalization.

High-performing dimensions. Across models, performance is consistently strong on dimensions related to normative and stylistic aspects of medical communication, including *Safety- and Compassion-Oriented Care*, *Structure and Emphasis in Presentation*, *Tone Adaptation in Clinical Contexts*, and *Translation of Medical Terminology into Lay Language*. Most models achieve near-saturated scores (often exceeding 90%) on these dimensions, indicating that contemporary LLMs are already proficient in clarity-oriented and form-driven medical communication.

Low-performing dimensions. In contrast, performance remains uniformly low on dimensions that require latent emotional reasoning and personalization, such as *Cultural Context Consideration*, *Individual Emotional Value*, *Understanding Emotional Attribution*, *Mixed Emotions*, *Emotional Transition*, and *Real-Time Emotional Response to Illness*. Average scores on these dimensions are typically below 30%, suggesting that current models struggle to infer implicit beliefs, evolving affective states, and patient-specific emotional meaning.

Implications. The observed gap indicates that while existing LLMs excel at surface-level emotional expression and communication structure, they remain limited in deeper emotional understanding and adaptive personalization. Notably, even the strongest models show only marginal improvements on low-performing dimensions, highlighting these competencies as key open challenges for emotionally competent medical AI systems.

5.4 Case Study

We present a pediatric case involving high fever and ankle swelling with suspected osteomyelitis to qualitatively illustrate differences captured by MedEQ. The scenario requires both accurate medical reasoning and emotionally appropriate communication.

High-scoring models provide structured explanations that combine clinical reasoning with explicit emotional acknowledgment and guidance, helping caregivers understand the situation while reducing anxiety. In contrast, lower-scoring responses focus primarily on differential diagnosis or factual listing, with limited emotional framing or personalization.

Notably, although GPT-5 demonstrates strong emotional consideration in general conversational settings, it exhibits a clear style shift in medical con-

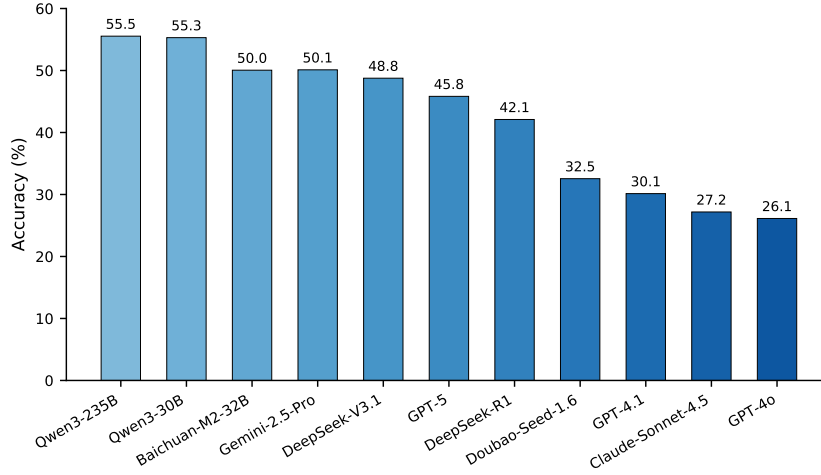


Figure 3: Performance comparison of state-of-the-art large language models on MedEQ. Scores are normalized rubric scores. Reasoning-oriented models tend to perform better than general-purpose models, although differences within the strongest model families can be small.

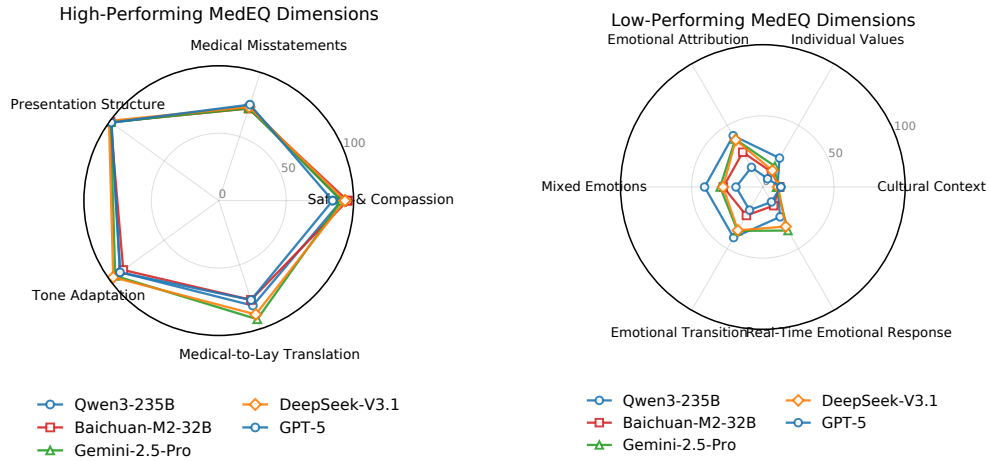


Figure 4: **Left:** High-performing MedEQ dimensions, primarily reflecting safety, compassion, and effective medical communication. **Right:** Low-performing MedEQ dimensions, highlighting challenges in nuanced emotional understanding and culturally grounded emotional reasoning.

texts. Its responses adopt a more formal and safety-oriented tone, emphasizing urgency and procedural guidance while reducing warm and approachable language. Full response examples are provided in Appendix F.

6 Rubric Validation via Reinforcement Learning

Beyond evaluation, we investigate whether the expert-validated rubrics can serve as effective reward signals for reinforcement learning (RL), thereby providing an independent validation of rubric quality. If the rubrics capture clinically meaningful emotional intelligence criteria, optimizing against them should improve emotional compe-

Table 2: Comparison of medical capability and emotional intelligence across representative models. Qwen3-32B (trained with rubrics) denotes the model fine-tuned via reinforcement learning using rubric-based reward signals.

Model	MedQA	MedMCQA	CMMLU (Med)	EQ-Bench
Qwen3-235B	91.98	56.44	90.03	48.8
Gemini-2.5-Pro	89.08	83.96	89.91	66.1
DeepSeek-R1	93.64	79.20	90.15	61.6
Qwen3-32B	85.47	72.27	89.18	42.42
Qwen3-32B (rubrics)	88.14	73.20	89.16	81.49

tence without degrading medical knowledge.

Design rationale. MedEQ is designed as a *pure evaluation benchmark*: the 413 expert-validated rubric instances constitute a held-out test set, not

a training resource. In this section, we repurpose the rubric scoring function as a reward signal to validate whether the rubrics encode learnable emotional intelligence criteria. Because the rubric items themselves serve as the reward during training, reporting MedEQ scores for the resulting model would conflate the optimization objective with the evaluation target. We therefore assess generalization on the independent EQ-Bench benchmark, which measures emotional intelligence through a different task format and data distribution, providing an uncontaminated signal of whether rubric-guided training transfers beyond the reward source.

Training procedure. We apply *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024) to Qwen3-32B using the rubric-based normalized score (Section 4.4; full formulation in Appendix D) as the reward signal $R(x, y)$. For each input x , a group of N candidate outputs is sampled from the current policy π_θ , and advantages are computed relative to the group $A_i = \frac{R_i - \mu_R}{\sigma_R}$, where μ_R and σ_R are the group mean and standard deviation. Because advantages are computed within each prompt group, GRPO is less sensitive to prompt-specific reward scaling and stochastic variation in the judge, making it suitable for noisy rubric-based rewards.

Results. As shown in Table 2, rubric-guided RL does not degrade medical performance. On core medical benchmarks including MedQA, MedMCQA, and CMMLU (Med), the rubric-trained model achieves comparable or slightly improved results relative to the base Qwen3-32B. In contrast, the same training procedure yields a substantial improvement in emotional intelligence: on EQ-Bench, Qwen3-32B trained with rubrics improves from 42.42 to 81.49, significantly exceeding the EQ performance of larger frontier models.

These results provide two-fold validation: (1) the expert-refined rubrics capture meaningful emotional intelligence criteria that can be optimized via RL, and (2) rubric-guided training improves emotional competence without sacrificing medical knowledge, confirming that clinical emotional intelligence and medical accuracy are not inherently in tension.

7 Conclusion

We introduced MedEQ, a benchmark for evaluating clinical emotional intelligence in patient-facing medical LLMs. Using expert-refined rubrics, it

assesses emotion recognition, empathetic communication, uncertainty handling, and affective safety. Current LLMs perform well on surface-level aspects (clear structure, safe wording, general compassion) but struggle with deeper emotional understanding, personalization, cultural context, mixed emotions, and illness-related distress. A proof-of-concept RL experiment showed that MedEQ-derived rubric signals can boost performance on an independent EQ benchmark without degrading standard medical QA. Overall, MedEQ offers a structured framework to measure and analyze emotional intelligence in medical LLMs.

8 Limitations

The benchmark contains 413 scenarios, which enables dense rubric coverage but remains smaller than broad open-domain dialogue datasets. Meanwhile, the rubric-based evaluation can favor explicit emotional acknowledgment, so models that communicate empathy in more concise or culturally implicit ways may be under-scored. The six-point scale has no neutral midpoint and may force borderline cases toward one side of the scale. We use it to distinguish complete absence from graded criterion satisfaction, but alternative scale designs should be examined in future work. Experts revised LLM-generated rubric drafts, and no independently human-authored rubric set was created for comparison. The initial drafts may therefore have anchored subsequent expert judgments. Although every retained rubric item underwent expert revision and validation, we did not include a fully human-authored control condition, so the extent of this automation bias remains unknown. Finally, the GRPO validation uses one base model and one external EI benchmark; whether the observed gains generalize to other model families, languages, training settings, and clinical contexts remains unknown.

Acknowledgments

Generative AI tools were used solely for language editing and polishing. All scientific content was reviewed and verified by the authors. This work was also supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant No. 2026ZD0558202).

References

- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimplouras, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *arXiv preprint arXiv:2505.08775*.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, and 1 others. 2025. [Medhelm: Holistic evaluation of large language models for medical tasks](#). *arXiv preprint arXiv:2505.23802*.
- Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.
- Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. [Emotionqueen: A benchmark for evaluating empathy of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176. Association for Computational Linguistics.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. 2013. [Effectiveness of empathy in general practice: a systematic review](#). *The British Journal of General Practice*, 63(606):e76–e84.
- Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Virginia Eatough and Jonathan A Smith. 2017. Interpretative phenomenological analysis. In Carla Willig and Wendy Stainton Rogers, editors, *The SAGE Handbook of Qualitative Research in Psychology*, 2 edition, pages 193–211. SAGE Publications Ltd London.
- Ronald M Epstein and Richard L Street. 2011. The values and value of patient-centered care. *The Annals of Family Medicine*, 9(2):100–103.
- Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M Schwartz, and Steven Woloshin. 2007. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96.
- Geoffrey H Gordon, Sandra K Joos, and Jennifer Byrne. 2000. Physician expressions of uncertainty during patient encounters. *Patient education and counseling*, 40(1):59–65.
- Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, and 1 others. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.
- Yuanyuan He, Yongsan Pan, Wei Li, Jiali You, Jiawen Deng, and Fuji Ren. 2025. Ecc: An emotion-cause conversation dataset for empathy response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6000–6017.
- Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K Arora, Foivos Tsimplouras, Preston Bowman, Michael Sharman, Chi Tong, Kevin Karthik, Arnav Dugar, Akshay Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. 2026. [Healthbench professional: Evaluating large language models on real clinician chats](#). *arXiv preprint arXiv:2604.27470*.
- He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. 2025a. Emobenchmark: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.
- He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. 2025b. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*.
- Arthur Kleinman. 1980. *Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry*, volume 3. Univ of California Press.
- Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.
- Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.
- Hao Liang, Linzhuang Sun, Jingxuan Wei, Xijie Huang, Linkun Sun, Bihui Yu, Conghui He, and Wentao Zhang. 2024. Synth-empathy: Towards high-quality synthetic empathy data. *arXiv preprint arXiv:2407.21669*.
- Zachary C Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.
- Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Guan-Ting Lin, and 1 others. 2025. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. *arXiv preprint arXiv:2508.17623*.

- Neil Mallinar, A Ali Heydari, Xin Liu, Anthony Z Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, and 1 others. 2025. A scalable framework for evaluating health language models. *arXiv preprint arXiv:2503.23339*.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Delip Rao and Chris Callison-Burch. 2026. [Aurubric: A unifying framework for rubric-based llm evaluation on non-verifiable tasks](#). *arXiv preprint arXiv:2603.00077*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.
- Donald A Redelmeier, Paul Rozin, and Daniel Kahneman. 1993. Understanding patients’ decisions: cognitive and emotional perspectives. *Jama*, 270(1):72–76.
- Helen Riess. 2017. The science of empathy. *Journal of patient experience*, 4(2):74–77.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Richard L Street Jr, Gregory Makoul, Neeraj K Arora, and Ronald M Epstein. 2009. How does communication heal? pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, 74(3):295–301.
- John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642:442–450.
- Xu Wang, Bo Wang, Yihong Tang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. Ecc: Synergizing emotion, cause and commonsense for empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5475–5485.
- Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023a. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.
- Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2025. Kardian-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. *arXiv preprint arXiv:2512.01282*.
- Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, and 1 others. 2025. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*.

A Model Details and Name Mapping

For readability, abbreviated model names are used throughout the main paper. This appendix provides the exact correspondence between abbreviated names and full model identifiers, along with approximate release dates and brief descriptions. All evaluations are conducted using the exact model versions listed below.

Qwen3-235B Full name: Qwen3-235B-A22B-Thinking-2507

A very large-scale reasoning-oriented model from the Qwen3 series, featuring explicit thinking mechanisms and a mixture-of-experts architecture, optimized for complex multi-step reasoning.

Qwen3-30B Full name: Qwen3-30B-A3B-Thinking-2507

A smaller reasoning-focused variant within the Qwen3 family, sharing the same architectural design as Qwen3-235B with reduced model capacity.

Baichuan-M2-32B Full name: Baichuan-M2-32B-Thinking

An open-source medical-oriented model designed to integrate domain knowledge with structured reasoning for clinical and biomedical tasks.

Gemini-2.5-Pro Full name: Gemini-2.5-Pro-With-Thinking

A proprietary large language model optimized for advanced reasoning and long-context understanding across complex domains, including medicine.

DeepSeek-V3.1 Full name: DeepSeek-V3.1-Terminus-Thinking

An updated reasoning-centric model from the DeepSeek series, incorporating architectural and training improvements over earlier versions.

GPT-5 Full name: GPT-5-2025-08-07-Thinking-High

A next-generation proprietary model with enhanced reasoning capabilities enabled via a high-level reasoning configuration.

DeepSeek-R1 Full name: DeepSeek-R1-0528

An earlier reasoning-oriented model in the DeepSeek family, with weaker long-horizon reasoning compared to later versions.

Doubao-Seed-1.6 Full name: Doubao-Seed-1.6-thinking-2025-07-15

A general-purpose large language model with optional reasoning modes, not specifically optimized for medical reasoning.

GPT-4.1 Full name: GPT-4.1-2025-04-14

An incremental update to the GPT-4 series focusing on improved instruction following and efficiency.

Claude-Sonnet-4.5 Full name: Claude-sonnet-4.5-20250929

A mid-sized proprietary model optimized for safety and conversational quality rather than domain-intensive reasoning.

GPT-4o Full name: GPT-4o-2024-11-20

A multimodal general-purpose model emphasizing speed and versatility over deep domain-specific reasoning.

B Category Definitions

Tables 3 and 4 detail the dimension and department taxonomies used in the coverage heatmap (Figure 2a).

B.1 Evaluation Dimensions

We organize the 21 evaluation dimensions into three categories that reflect complementary aspects of affective competence in clinical communication:

Empathic Communication Competence covers the foundational skills required for effective physician–patient interaction, including the ability to respond to disease-related emotions in real time, guide clinical inquiries, affirm patient behaviors, translate medical jargon into accessible language, adapt register and tone to the clinical context, and structure information with emotional sensitivity throughout the encounter.

Clinical Emotional Literacy captures a clinician’s capacity to perceive, interpret, and reason about complex affective states. This category encompasses recognizing emotional transitions and mixed affect, anticipating counterintuitive emotional outcomes, accounting for cultural backgrounds and personal values, analyzing role-driven influences, decoding implicit signals, understanding emotional attributions, navigating miscommunication contexts, inferring atypical pathology-related affect, and correcting medical misconceptions.

Affective Safety Praxis addresses the action-oriented competencies that translate emotional understanding into safe, patient-centered care, including self-regulation in emotionally charged clinical contexts, adherence to safety-centric care protocols, and the design of personalized intervention plans.

B.2 Clinical Department Groups

The 51 clinical departments appearing in our dataset are consolidated into 12 groups based on medical specialty affinity (Table 4). Core clinical disciplines—*Internal Medicine*, *Surgery*, and *Pediatrics*—each encompass multiple sub-specialties and account for the largest share of scenarios. Population-specific departments are grouped under *OB/GYN & Reproductive*. *Oncology*, *Infectious Disease*, and *Psychiatry* are kept as separate groups due to their distinct emotional dynamics. *TCM* (Traditional Chinese Medicine) is isolated to reflect its unique diagnostic and communicative paradigms. Departments involving the head, neck, and sensory organs are combined into *Head, Neck & Sensory*. Acute-care settings form the *Emergency & Critical Care* group. Non-clinical or ancillary departments such as laboratory, imaging, and radiology are placed under *Diagnostics & Support*, while remaining specialties are collected in *Other Specialties*.

C MedEQ Rubric Taxonomy

Table 5 summarizes the 21 foundational dimensions used in the MedEQ rubric.

The MedEQ rubric is designed to operationalize emotionally competent medical communication in a manner that is both interpretable and amenable to automated evaluation. The rubric contains 21 primary evaluation dimensions. Macro-competencies and thematic groups provide an organizational view of these dimensions rather than additional nested scoring levels.

Macro-Competency Structure. For analysis, rubric dimensions are grouped into three macro-competencies: *Empathic Communication*, *Clinical Emotional Literacy*, and *Affective Safety*. These categories correspond to distinct but interrelated requirements for emotionally appropriate clinical dialogue.

Empathic Communication. Empathic Communication concerns the expression of emotional awareness while maintaining clarity and clinical relevance. Prior work in health communication shows that empathic language improves patient experience even when controlling for informational content (Riess, 2017; Derksen et al., 2013). In this rubric, empathy is treated as a composite capacity that includes affective acknowledgment, register adaptation, information structuring, and translation of technical content into accessible language. From a cognitive load perspective, emotional distress can impair information processing, making readability and organization integral components of effective communication rather than optional stylistic features (Sweller, 1988; Street Jr et al., 2009).

Clinical Emotional Literacy. Clinical Emotional Literacy captures the inferential aspects of emotional competence, including the ability to reason about implicit, mixed, or dynamically evolving emotional states. Clinical psychology research indicates that patient emotions often cannot be inferred from surface sentiment alone, particularly in cases involving serious or chronic conditions (Lazarus, 1991; Eatough and Smith, 2017). Accordingly, the rubric includes criteria related to emotional transitions, belief-driven interpretation, perspective-taking, and the resolution of common medical misconceptions, which are known contributors to emotional distress and communication breakdowns (Kleinman, 1980; Epstein and Street, 2011).

Table 3: Evaluation dimension taxonomy. Each of the 21 fine-grained dimensions is grouped into one of three overarching categories.

Category	Dimension
Empathic Communication Competence	Disease-Related Emotional Responsiveness
	Clinical Inquiry Navigation
	Therapeutic Affirmation
	Jargon-to-Lay Translation
	Clinical Register Adaptation
	Content Structuring
	Emotionally-Integrated Interaction
Clinical Emotional Literacy	Emotional Transition
	Mixed Affect States
	Counterintuitive Emotional Outcomes
	Cultural Contextualization
	Personal Value Recognition
	Role-Driven Affect Analysis
	Implicit Signal Decoding
	Attributional Reasoning
	Clinical Miscommunication Contexts
	Atypical Pathology-Affect Inference
	Medical Misconception Resolution
Affective Safety Praxis	Self-Regulation in Clinical Contexts
	Safety-Centric Care Protocols
	Personalized Intervention Design

Affective Safety. Affective Safety provides a normative constraint on emotional engagement. While emotional responsiveness is generally beneficial, uncalibrated reassurance or excessive affective alignment can inadvertently reinforce maladaptive beliefs or obscure medical uncertainty (Redelmeier et al., 1993; Gordon et al., 2000). This category evaluates whether emotional language is appropriately regulated, aligned with safety-oriented care principles, and consistent with the clinician’s role. Its inclusion is informed by work on risk communication emphasizing the balance between emotional support, accuracy, and caution in high-stakes clinical contexts (Gigerenzer et al., 2007).

tant in medical applications where understanding the source of a failure is often as important as measuring its frequency.

Fine-Grained Criteria and Diagnostic Use. Across all macro-competencies, each rubric dimension is decomposed into minimal, observable criteria. Broad labels such as “empathetic” or “emotionally aware” conflate distinct behaviors and failure modes, limiting their utility for evaluation and model improvement. Both measurement theory and interpretability research emphasize the importance of fine-grained, behavior-linked criteria for attribution, error analysis, and system comparison (Doshi-Velez and Kim, 2017; Lipton, 2018).

Overall, the rubric is intended as a diagnostic framework rather than a purely aggregative scoring tool. This organization enables analysis at multiple levels of abstraction, which is particularly impor-

Table 4: Clinical department taxonomy. The 51 fine-grained departments are organized into 12 groups along the horizontal axis of the heatmap.

Department Group	#	Included Departments
Pediatrics	6	Pediatric Medicine, Pediatric Surgery, Pediatric ENT, General Pediatrics, Child Healthcare, Neonatology
OB/GYN & Reproductive	5	Obstetrics, OB/GYN, Women’s Healthcare, Gynecology, Reproductive Medicine
Internal Medicine	6	Endocrinology, Pulmonology, Cardiology, Gastroenterology, Neurology, Hematology
Surgery	10	General Surgery, Urology, Thyroid & Breast Surgery, Neurosurgery, Hepatobiliary Surgery, GI Surgery, Thoracic Surgery, Vascular Surgery, Orthopedics, Anorectal Surgery
Oncology	2	Medical Oncology, Oncology
TCM	3	TCM Internal Medicine, TCM OB/GYN, TCM General Clinic
Head, Neck & Sensory	4	ENT & Head/Neck Surgery, Ophthalmology, Stomatology, Dermatology
Emergency & Critical Care	2	Emergency Medicine, Critical Care
Infectious Disease	2	Infectious Disease, Infection
Psychiatry	1	Psychiatry
Diagnostics & Support	3	Clinical Laboratory, Medical Imaging, Radiology
Other Specialties	7	Rehabilitation, Nutrition, Pain Medicine, Allergy, Laser Therapy, Multidisciplinary Clinic, Other

D Scoring Details

This section provides the full formal specification of the rubric-based scoring procedure summarized in Section 4.4.

Rubric-Weighted Scoring. Given a rubric with K criteria, the judge assigns a discrete score

$$s_k(y) \in \{0, 1, 2, 3, 4, 5\}$$

to each criterion k . A score of 0 indicates that the criterion is absent or not applicable, while scores from 1 to 5 indicate increasing levels of criterion satisfaction. We use uniform weights, with $w_k = 1$ for all criteria. The weighted score is

$$S_{\text{raw}}(y) = \sum_{k=1}^K w_k s_k(y),$$

and the maximum possible score is

$$S_{\text{max}} = \sum_{k=1}^K 5w_k.$$

We normalize the score as

$$S_{\text{norm}}(y) = \frac{S_{\text{raw}}(y)}{S_{\text{max}}} \in [0, 1].$$

All reported MedEQ scores are normalized to percentages by multiplying S_{norm} by 100 and averaging across instances.

Hard Constraints. Let $\{c_m(y)\}_{m=1}^M$ denote binary indicators for hard constraints such as format validity, repetition limits, or language requirements. An output is considered invalid if any constraint is violated:

$$\mathbb{I}_{\text{invalid}}(y) = 1 - \prod_{m=1}^M c_m(y).$$

Invalid outputs receive a fixed penalty score $R_{\text{neg}} < 0$.

Final Score. The overall evaluation score for an output is

$$R(x, y) = \begin{cases} R_{\text{neg}}, & \text{if } \mathbb{I}_{\text{invalid}}(y) = 1, \\ S_{\text{norm}}(y), & \text{otherwise.} \end{cases}$$

Judge Selection and Reliability. We use GPT-4.1 as the primary judge because it combines strong expert agreement (F1 = 0.855, Kappa = 0.601), low repeated-run variance, high cross-judge consistency, and practical evaluation efficiency. Prior work suggests that LLM-based judges can exhibit moderate to strong agreement with human evaluators for rubric-based judgments, while offering greater scalability. However, such judges may display biases toward fluent or verbose outputs and sensitivity to prompt design. Accordingly, we treat the resulting scores as noisy proxies for human judgment rather than ground-truth signals, and combine them with normalization and hard constraints to mitigate known failure modes.

Table 5: Organizational view of the MedEQ rubric. The 21 evaluation dimensions are presented with their macro-competency categories and thematic groupings; these groupings support presentation and analysis rather than define three nested scoring levels.

Macro-competency	Thematic Group	Evaluation Dimension
Empathic Communication Competence	Disease-Related Emotional Responsiveness	Guided Questioning in Clinical Communication Positive Reinforcement of Health Behaviors Safety- and Compassion-Oriented Medical Care Real-Time Emotional Response to Illness
	Readability Optimization	Translation of Medical Terminology into Lay Language Tone Adaptation in Clinical Context Structure and Emphasis in Presentation
Clinical Emotional Literacy	Pathology-Linked Emotional Complexity	Emotional Transition Mixed Emotions Unexpected Emotional Outcomes Emotionally Continuous Experience
	Illness Belief Mapping	Cultural Context Consideration Individual Emotional Value Role Influence
	Affective Cue Detection	Suggestive Cues Understanding Emotional Attribution
	Perspective-Taking	Medical Communication Misstatement Scenarios Atypical Emotional Reasoning in Illness Misconceptions About Illness Content Structure
Affective Safety Praxis	Self-Regulation in Clinical Contexts	Personal Emotion and Behavior Regulation
	Personalized Intervention Design	Personalized Care Plan

E Human-LLM Agreement Evaluation

This evaluation is separate from the 30-expert rubric-construction process in Section 4.3. We selected 100 evaluation instances using stratified sampling across score ranges and rubric dimensions. Twenty domain experts with clinical or psychology backgrounds and 20 general-public annotators independently applied the finalized rubrics, producing 3,632 rubric-level annotations.

Each “Clinical Question” in Appendix H represents the complete de-identified scenario presented to the model, preserving the information available in the original consultation query without adding inferred contextual details such as the identity or role of the questioner.

A well-designed rubric should be *executable*: given the rubric text alone, both trained human professionals and capable language models should arrive at consistent judgments without relying on implicit knowledge or ad-hoc interpretation. We evaluate this executability property along three axes: (1) alignment between LLM judges and domain-expert annotations, (2) the contrast between domain-expert and general-public alignment,

and (3) inter-model consensus among diverse judge LLMs.

Together, these analyses demonstrate that the MedEQ rubrics are highly executable for both domain professionals and state-of-the-art language models: experts and top-performing judges converge on the same assessments, while general-public annotators show substantially lower agreement, confirming that the rubrics encode domain-specific criteria rather than surface-level heuristics.

Metrics. We report F1 score, Precision, Recall, Cohen’s Kappa, and Matthews Correlation Coefficient (MCC). Kappa and MCC explicitly correct for chance agreement and class imbalance, providing robust measures of true concordance.

We first examine alignment between model judgments and expert annotations. Table 6 reports the F1 score of each model against expert labels. All evaluated models achieve high F1 scores, indicating strong overall agreement with expert judgment.

Tables 7 and 8 further report Precision and Recall with respect to expert annotations, revealing that while models exhibit different evaluation tendencies, all remain consistent and well aligned.

Table 6: F1 score between model judgments and expert annotations.

Model	F1
kimi_k2	0.862
ds_v31	0.861
gpt_4.1	0.855
gemini_2.5_pro	0.853
gemini_3_pro_preview	0.842
qwen3_235b_a22b	0.823
ds_r1	0.811

Table 7: Precision with respect to expert annotations.

Model	Precision
ds_v31	0.847
ds_r1	0.841
gemini_3_pro_preview	0.832
kimi_k2	0.819
gpt_4.1	0.802
gemini_2.5_pro	0.800
qwen3_235b_a22b	0.775

Cohen’s Kappa and MCC, shown in Table 9, indicate moderate to strong agreement beyond chance for all models.

We additionally evaluate alignment with general-public annotations. As shown in Table 10, agreement scores decrease uniformly across all models: domain-expert-aligned Kappa values of 0.517–0.646 drop to 0.341–0.419 against general-public labels. This gap is consistent across all seven judge models and indicates that the rubrics encode domain-specific clinical criteria that trained professionals and capable LLMs can reliably apply, whereas general-public annotators interpret the same criteria with substantially greater subjectivity.

To assess stability, each model is evaluated three times on identical rubric instances. Table 11 reports the mean standard deviation and variance of the resulting scores, showing that top-performing judges exhibit highly stable behavior.

Inter-model consensus. Beyond human–model alignment, we examine whether different judge LLMs reach consistent conclusions when applying the same rubrics independently. Table 12 reports pairwise Cohen’s Kappa for the highest-consensus cluster identified among the seven evaluated judge models. Kimi-K2, DeepSeek-V3.1, and GPT-4.1 form a tightly aligned group with pair-

Table 8: Recall with respect to expert annotations.

Model	Recall
gpt_4.1	0.916
gemini_2.5_pro	0.915
kimi_k2	0.909
qwen3_235b_a22b	0.878
ds_v31	0.876
gemini_3_pro_preview	0.852
ds_r1	0.783

Table 9: Cohen’s Kappa and MCC between model judgments and expert annotations.

Model	Kappa	MCC
ds_v31	0.646	0.647
kimi_k2	0.629	0.635
gpt_4.1	0.601	0.612
gemini_3_pro_preview	0.601	0.601
gemini_2.5_pro	0.596	0.606
ds_r1	0.554	0.556
qwen3_235b_a22b	0.517	0.525

wise Kappa values of 0.736–0.784, substantially exceeding their agreement with other models. Notably, these three models are also the top-ranked judges in expert alignment (Table 9), confirming that their consensus reflects convergence toward expert-calibrated judgment rather than shared bias.

Finally, Table 13 reports approximate wall-clock time for a full evaluation run. GPT-4.1 achieves a favorable balance between evaluation quality, stability, and efficiency.

Overall, across human alignment, intra-model stability, and inter-model consensus analyses, we find consistent evidence that the MedEQ rubrics are highly executable: trained professionals and state-of-the-art LLMs converge on the same assessments (domain-expert Kappa 0.517–0.646; top inter-model Kappa 0.736–0.784), while general-public annotators show substantially lower agreement (Kappa 0.341–0.419). This gap between domain experts and general-public annotators confirms that the rubrics encode clinically meaningful criteria rather than surface heuristics, and that their executability scales to automated evaluation. In particular, GPT-4.1 lies within the highest-consensus group of judges, aligns closely with expert judgment, and exhibits low variance under repeated evaluations, making it a well-justified choice for

Table 10: Cohen’s Kappa and MCC between model judgments and general-public annotations.

Model	Kappa	MCC
kimi_k2	0.419	0.423
ds_v31	0.413	0.413
gpt_4.1	0.410	0.417
gemini_2.5_pro	0.380	0.386
gemini_3_pro_preview	0.378	0.378
ds_r1	0.362	0.363
qwen3_235b_a22b	0.341	0.345

Table 11: Intra-model variance across three repeated evaluations.

Model	Std. Dev.	Variance
ds_r1	0.101	0.058
ds_v31	0.032	0.018
gemini_2.5_pro	0.042	0.024
gemini_3_pro_preview	0.0003	0.00018
gpt_4.1	0.016	0.009
kimi_k2	0.026	0.015
qwen3_235b_a22b	0.101	0.058

automatic evaluation.

Table 12: Inter-model Cohen’s Kappa. The highest-consensus cluster (Kimi-K2, DS-V3.1, GPT-4.1) achieves pairwise Kappa 0.736–0.784.

	DS-V3.1	GPT-4.1	Kimi-K2
DS-V3.1	—	0.736	0.779
GPT-4.1	0.736	—	0.784
Kimi-K2	0.779	0.784	—

Table 13: Approximate wall-clock time required for full evaluation.

Model	Time
ds_r1	5–5.5 h
ds_v31	1.5–2 h
gemini_2.5_pro	3–3.5 h
gemini_3_pro_preview	3.5–4 h
gpt_4.1	1–1.5 h
kimi_k2	< 1 h
qwen3_235b_a22b	2.5–3 h

F Case Study Responses

This appendix provides the full responses underlying the qualitative analysis in Section 5.4. The high- and low-score labels refer to MedEQ rubric scores rather than clinical correctness, and the color highlights are explanatory annotations that do not affect scoring. The GPT-5 response illustrates the safety-oriented style shift discussed in the main text.

Table 14: Translated case study responses for a pediatric patient with high fever and ankle swelling and suspected osteomyelitis. **Blue text** highlights empathic and emotionally supportive language; **red text** marks overly safety-strict or urgency-focused language that lacks emotional warmth. The low-score response contains no emotionally supportive language and is therefore shown without color annotation.

Response Type	Response Content (English Translation)
Clinical Question	A child presents with high fever accompanied by ankle joint swelling, and the clinician suspects osteomyelitis.
High-Score Response	Seeing a child with persistent high fever accompanied by ankle joint swelling is indeed a situation that requires serious attention. As a parent, the anxiety and concern at this moment are completely understandable. High fever in children combined with swelling, pain, and restricted movement of a single joint—especially the ankle—is a classic presentation that necessitates urgent evaluation for osteomyelitis. At the same time, it is also important to differentiate this condition from several other critical diseases. The following is a detailed analysis to help you understand the clinician’s reasoning and what may happen next. Acute osteomyelitis in children often has a sudden onset and is commonly associated with high fever (usually above 38.5°C), chills, and other systemic signs of infection. In pediatric patients, the metaphysis of long bones is the most frequently affected site. The distal tibia and fibula near the ankle joint are among the most common locations for metaphyseal infection. The affected area, such as the ankle region, lower leg, or area above the heel, typically exhibits severe and persistent pain. Children may refuse to move the affected limb or bear weight, and even slight touch or passive movement can trigger intense pain and crying. Marked swelling, skin redness, and increased local temperature may be observed in the infected area and surrounding soft tissue. Ankle joint mobility is often significantly restricted.
Low-Score response	When a child presents with high fever accompanied by ankle joint swelling, the following conditions should be considered. Septic arthritis is characterized by infection within the joint cavity. Typical manifestations include redness, swelling, warmth, and pain of the joint, with marked limitation of movement, especially severe pain during passive motion. High fever and systemic signs of infection are common. The hip and knee joints are more frequently affected. Acute osteomyelitis is characterized by infection of the bone marrow and bone tissue. Clinical manifestations include bone pain and localized tenderness. Swelling of adjacent joints may occur; however, joint mobility is often relatively preserved.
GPT-5 Response	A child presenting with high fever and ankle joint swelling requires immediate medical attention and emergency evaluation. This presentation suggests possible acute osteomyelitis or septic arthritis, both of which are pediatric orthopedic emergencies. Delayed treatment may result in joint destruction, impaired bone growth, or systemic infection. Immediate actions include proceeding to the emergency department or following the physician’s recommendation for hospitalization. If the child is unable to walk or bear weight, has obvious redness, swelling, warmth of the ankle, persistent high fever (at or above 39°C), poor mental status, or chills, urgent evaluation is essential. During transport, antipyretic and analgesic treatment may be administered, such as acetaminophen at 10–15 mg/kg every 6–8 hours or ibuprofen at 10 mg/kg every 8 hours, without exceeding recommended doses. Ibuprofen should be avoided in children with dehydration, renal disease, or gastrointestinal conditions. Aspirin should never be given to children. The child should rest, movement of the affected limb should be restricted, and hydration should be encouraged with small, frequent fluid intake. In the hospital, evaluation typically includes blood tests such as complete blood count, C-reactive protein, erythrocyte sedimentation rate, and possibly procalcitonin, with blood cultures obtained before antibiotic administration whenever possible. Imaging studies may include radiographs, which are often normal early on, ultrasound to detect joint effusion, and MRI, which is the most sensitive modality for assessing the extent of osteomyelitis and detecting abscess formation. In some cases, joint aspiration or bone or soft tissue drainage may be required for diagnostic confirmation and decompression, with samples sent for culture. Empiric intravenous antibiotics should be initiated promptly and adjusted based on culture results. Coverage typically includes <i>Staphylococcus aureus</i>, including methicillin-resistant strains. In preschool-aged children, <i>Kingella kingae</i> should also be considered. In children with sickle cell disease, <i>Salmonella</i> species must be considered.

G Bilingual Dataset Release

MedEQ is annotated and released in both Chinese and English. All rubric items, scenario descriptions, and evaluation criteria are provided in both languages. The results reported in this paper are based on the English version of the dataset. Both versions will be made publicly available upon publication. The dataset will be released under the CC BY-NC 4.0 license for non-commercial research use.

H Rubric Case Examples

To illustrate the depth and clinical sophistication of MedEQ rubrics, we present two representative cases with selected rubric items. These examples demonstrate how the rubric system operationalizes nuanced emotional intelligence requirements that go far beyond surface-level empathy detection. The examples below present rubric criteria without pre-assigned point values. During evaluation, every criterion is independently scored on the same 0–5 scale and assigned unit weight ($w_k = 1$).

H.1 Case 1: End-of-Life Caregiving

Clinical query. *“For an elderly person with stage 4 heart failure and 90% lung cancer, how long can they live?”*

This scenario involves a family caregiver confronting terminal illness in an elderly relative. The question “how long can they live?” appears to request a prognosis, but its emotional subtext encompasses anticipatory grief, caregiver guilt, helplessness, and the need for actionable guidance. The rubric must capture not only empathic acknowledgment but also the ability to navigate implicit meanings, conflicting emotions, role-specific pressures, and the ethical boundary between honesty and emotional safety. The following selected rubric items highlight these requirements:

1. **Mixed Emotions.** Deeply understands the asker’s conflicting psychological state—holding onto “hope for a miracle” while simultaneously preparing for the worst. The response skillfully addresses this emotional complexity, neither offering false hope nor completely extinguishing their emotional anchor, but instead guiding them to focus energy on more meaningful actions.
2. **Suggestive Cues.** Accurately interprets the underlying meanings behind the question “How long can they live?”—such as “How should I prepare?”, “How much time do I have left?”, or “What can I do to ensure their final days are as pain-free as possible?” The core of the response centers on these implicit concerns rather than fixating on the “timeline” itself.
3. **Understanding Emotional Attribution.** Anticipates the profound harm that a direct, specific answer (even a range like “3–6 months”)

could cause, and clearly, gently explains why doctors cannot and should not deliver a definitive “death sentence.” At the same time, skillfully shifts the focus from “length of life” to “quality of life.”

4. **Unexpected Emotional Outcomes.** Family members may repeatedly blame themselves—“Did I fail to detect the illness earlier?” or “Should I have pushed harder for the elder to quit smoking?”—even when these factors are not directly linked to the illness. Over long-term caregiving, family members constantly absorb the patient’s negative emotions, compounded by their own worries, making them prone to “emotional exhaustion.” The response advises caregivers to allow themselves to “occasionally show vulnerability,” such as confiding in someone trusted, to avoid bearing all the pressure alone.
5. **Role Influence.** Accurately empathizes with the role pressure of the asker as a “child” or “younger family member.” The response reflects an understanding of cultural expectations of filial duty, feelings of helplessness in the face of a loved one’s decline, and possible anticipatory grief and guilt associated with “wanting to care for one’s parent, but the parent is no longer there.”
6. **Emotionally Continuous Experience.** Maintains a consistent tone of care and empathy from beginning to end. At the conclusion, offers emotional support and encouragement again—such as “This journey is tough; please remember to take care of yourself—you’re not facing it alone”—creating a complete emotional loop.
7. **Personalized Care Plan.** While offering general recommendations (e.g., focusing on palliative care), encourages the user to explore the unique, personalized needs of this particular elder. For example: “Has the elder recently mentioned wanting to see someone? Is there a particular food they’d like to eat? Any unfinished wishes?” This ensures care is truly centered on one specific individual.
8. **Safety- and Compassion-Oriented Care.** All recommendations strictly uphold “safety first” and “compassionate care.” It clearly states that treatment decisions must be made

jointly with the medical team and aligns the goal with palliative care philosophy—in the final stage of life, the focus should be on pain control, maintaining dignity, and fulfilling final wishes, rather than extending life at all costs.

H.2 Case 2: Adolescent Hygiene in a Boarding School

Clinical query. *“As a boarding student, there’s no place to shower at school, so I use baby wet wipes to clean my private area before going to bed at night.”*

This scenario involves an adolescent navigating intimate hygiene under constrained living conditions. The emotional complexity lies in the intersection of potential embarrassment, limited autonomy, environmental constraints, and a common misconception (“baby products are always safe for sensitive areas”). The rubric requires sensitivity to cultural and situational context, recognition of the student’s resourcefulness despite limited options, and correction of misconceptions without blame. The following selected rubric items illustrate these requirements:

1. **Cultural Context Consideration.** Incorporates cultural and environmental factors common among boarding students, such as dormitory living and limited facilities. Reflects this understanding in responses, for example: “In a shared school environment, living conditions may be limited, and hygiene can present real challenges.”
2. **Individual Emotional Value.** Emphasizes the importance of private area health to a boarding student’s personal quality of life and academic performance. For example: “Maintaining intimate health helps you feel more comfortable in daily life and studies, reduces distractions from physical discomfort, and allows you to better enjoy campus life.”
3. **Role Influence.** Considers the student role of boarding students, and tailors advice to fit their academic and daily routines. Recommends simple, time-efficient cleaning methods, for example: “Using warm water to clean is convenient and won’t take up much of your study time.”
4. **Mixed Emotions.** Anticipates and addresses parents’ “unexpected emotions.” Understands

that parents might feel their concern is being dismissed if the advice seems “too simple” (e.g., just recommending water), and explains accordingly: “Although it sounds simple, water is actually the gentlest and safest option. . .”

5. **Misconceptions About Illness.** Gently anticipates and corrects potential misconceptions without blame. For instance, if a parent assumes “baby products are always safest,” the response might explain: “While baby wipes are gentle, they are designed for different purposes than intimate care. . .” rather than bluntly stating, “You can’t use them.”
6. **Emotional Transition.** Transforms potential anxiety of boarding students into a sense of reassurance. For example: “Although the current cleaning method may have some drawbacks, we can better protect intimate health by making a few simple adjustments—there’s no need to feel overly anxious.”
7. **Understanding Emotional Attribution.** Accurately understands the root cause of the boarding student’s anxiety: “No access to showers at school leads to improper cleaning, which in turn causes health concerns.” For example: “I understand your anxiety stems from not having shower access at school, forcing you to use baby wipes on sensitive areas, and worrying this might affect health.”
8. **Suggestive Cues.** Accurately interprets the underlying sense of helplessness behind a brief question. The response not only answers “Can I use it?” but proactively offers alternatives (“What to use instead if not?”), long-term solutions, and fills in the gaps left by the parent’s anxiety, addressing unasked questions.