

# Reinforcing Traceable Medical Reasoning with Critical Evidence Graph

Linjie Mu<sup>1,2,\*</sup>   Yannian Gu<sup>1,\*</sup>   Zhongzhen Huang<sup>1</sup>   Yakun Zhu<sup>1,2</sup>  
Shaoting Zhang<sup>1,†</sup>   Xiaofan Zhang<sup>1,2,†</sup>  
<sup>1</sup> Shanghai Jiao Tong University   <sup>2</sup> Shanghai Innovation Institute

## Abstract

Large language models (LLMs) with reasoning capabilities have shown impressive performance in clinical applications, yet the readability and traceability of their reasoning processes remain limited, as reasoning quality is often overlooked during training. To address this, we propose MedCEG, a framework that augments medical language models with clinically traceable reasoning by explicitly supervising the reasoning process through a critical evidence graph (CEG). Specifically, we curate challenging clinical cases and algorithmically construct a CEG for each sample via multi-model consensus as a high-quality, traceable reasoning pathway. The CEG reformulates the reasoning process for supervised fine-tuning and underpins a traceable reasoning procedure reward that evaluates generated reasoning against the CEG through deterministic graph operations measuring semantic matching, triplet recall and connected-component analysis. Experiments show that MedCEG surpasses existing methods among open-source medical models of comparable scale while producing clinically readable and traceable reasoning chains. Code and models are available at <https://github.com/LinjieMu/MedCEG>.

## 1 Introduction

Large Language Models (LLMs) with reasoning capabilities have demonstrated notable advances in various domains (Guo et al., 2025; Yang et al., 2025), with medical applications showing particular promise (Singhal et al., 2025; Tu et al., 2025). These models have achieved strong performance on complex medical tasks, including diagnostic reasoning (Liu et al., 2025b), treatment planning (Goh et al., 2025) and clinical question answering (Jin et al., 2021; Hager et al., 2024).

Despite these advances, a diagnostic outcome or treatment plan alone is rarely actionable in clinical practice. Physicians need to retrace the underlying reasoning process, from initial symptoms through differential diagnoses to the final assessment, to verify that the conclusion is well-grounded before incorporating it into clinical decision-making (Liévin et al., 2024; Van Veen et al., 2024). A transparent, step-by-step reasoning process therefore serves not only as a decision aid but also as a critical safeguard, enabling clinicians to verify whether a model has considered the relevant evidence, followed valid inferential steps and ruled out plausible alternatives (Singhal et al., 2025).

However, existing methods for enhancing medical reasoning in LLMs largely overlook the quality of the reasoning process. A prominent line of work applies reinforcement learning with outcome-oriented rewards to improve diagnostic accuracy (Liu et al., 2025a; Zhang et al., 2025), achieving notable gains on medical question-answering benchmarks. Yet these methods evaluate only the correctness of the final answer while leaving the intermediate reasoning process entirely unsupervised. When training is driven solely by answer accuracy, models may arrive at correct conclusions through clinically invalid reasoning paths. As illustrated in Figure 1, a model may leap directly to a diagnosis such as “chemical burn” while completely bypassing the evaluation of key pathological findings revealed by endoscopy.

Another line of work recognizes the importance of reasoning structure and leverages external knowledge or system to reconstruct reasoning pathways for supervised fine-tuning (Wu et al., 2025a; Sun et al., 2025). While this approach improves the factual grounding of reasoning chains, its supervision of the reasoning process remains insufficient. The model is trained to imitate reference reasoning traces directly through static pattern matching, without any mechanism to trace back through the

\*Co-first authors

†Corresponding authors

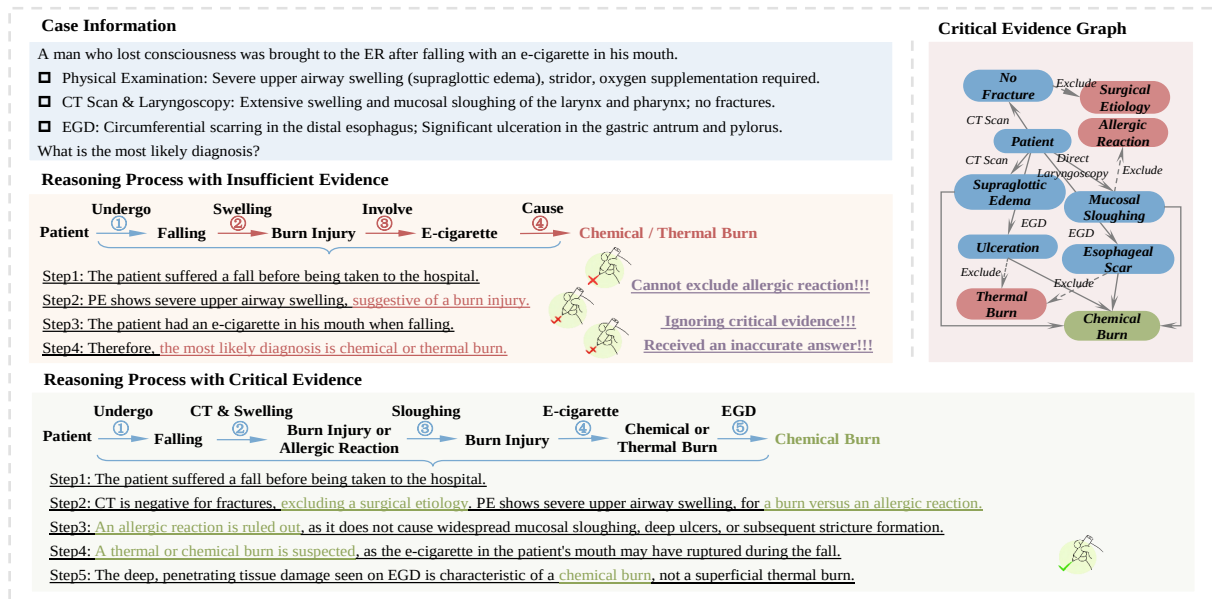


Figure 1: Comparison between a flawed, superficial reasoning process and a structured, evidence-driven pathway guided by the Critical Evidence Graph (CEG). The top process shows a flawed approach leading to a hasty conclusion. The bottom process, guided by the CEG, follows a systematic pathway to a more precise diagnosis.

generated process and provide reward signals that reinforce adherence to clinically valid inferential steps. Consequently, the model lacks an explicit, ongoing incentive to maintain reasoning quality, as the supervision never actively verifies whether the generated chain faithfully covers the necessary evidence or follows correct causal pathways.

Building on these observations, we propose a graph-based alignment framework that supervises the reasoning process itself, rather than merely the final answer. By converting clinical narratives into traceable evidence graphs, we construct a structured and explicit representation of the reasoning process that serves as the foundation for process-level reward computation. Our approach begins by algorithmically constructing instance-specific *Evidence Graphs* (EGs) through a multi-model consensus mechanism. Specifically, triplets are independently extracted by three frontier models from distinct providers and only those agreed upon by at least two models are retained, thereby suppressing model-specific noise and mitigating systematic biases that any single model might introduce. From each EG, we then extract a *Critical Evidence Graph* (CEG) via backward traversal and transitive reduction, capturing essential clinical entities and their causal relationships while representing the minimal, logically necessary backbone of the argument.

The EG and its refined CEG serve distinct pur-

poses in our training pipeline. First, the EG is linearized back into a coherent textual sequence for cold-start supervised fine-tuning (Guo et al., 2025), compelling the model to generate explanations that explicitly articulate the formal, step-by-step logical trajectory encoded in the graph. Second, the CEG provides a direct reward signal for a subsequent reinforcement learning phase, where the reward is computed entirely via rule-based graph operations, including semantic matching, triplet recall and connected-component analysis, rather than LLM-based scoring. This design ensures that reinforcement targets the most critical inferential steps and trains the model to prioritize traceable causal pathways essential for sound clinical reasoning.

We evaluate our approach on six medical benchmarks, including both in-domain and out-of-domain settings, and conduct comprehensive ablation studies alongside human expert evaluation by qualified medical professionals. Our contributions are summarized as follows.

- We construct a dataset of ten thousand clinical cases with evidence graphs built through multi-model consensus, explicitly modeling medical entities, relationships and reasoning pathways.
- We design a CEG-based reward mechanism that directly supervises the reasoning process, producing traceable and readable reasoning chains.
- We conduct extensive experiments on six med-

ical benchmarks, demonstrating significant improvements in both answering accuracy and reasoning quality, validated by controlled ablations and human-expert-model agreement analysis.

## 2 Related Work

### 2.1 RL Incentives for LLM Reasoning

Reinforcement learning has become a central paradigm for improving reasoning capability in LLMs (Yuan et al., 2024; Plaat et al., 2024). The most common formulation rewards only the correctness of the final answer (Guo et al., 2025), an approach that has yielded notable gains on reasoning benchmarks but provides no supervision over intermediate steps. Because the reward signal is sparse and terminal, models can arrive at correct conclusions through logically flawed or clinically invalid pathways. This limitation is especially concerning in medical settings, where the validity of each constituent reasoning step is of paramount importance (Liu et al., 2025a; Zhang et al., 2025).

### 2.2 Traceable Medical LLM Reasoning

Establishing traceable reasoning in medical LLMs necessitates transparency and logical coherence throughout the decision-making process. Chain-of-Thought (CoT, Wei et al., 2022) prompting offers a foundational solution by eliciting structured reasoning pathways, and systems such as MedReason (Wu et al., 2025a) further ground these chains in authoritative knowledge graphs (Chandak et al., 2023) to render model logic auditable. However, exposing models to static exemplars via supervised fine-tuning alone proves insufficient for instilling robust reasoning behavior. More recent efforts turn to reinforcement learning with process reward models to provide step-level supervision (Huang et al., 2025), as exemplified by Huatuo-o1 (Chen et al., 2024) and MedS<sup>3</sup> (Jiang et al., 2026). While effective, these approaches typically incur substantial annotation overhead or computational cost and remain susceptible to reward hacking, underscoring the need for more efficient and faithful strategies.

## 3 Data Preparation

We curate a clinical rationale corpus that combines challenging clinical questions with structured reasoning representations for developing rigorous reasoning capabilities. Detailed construction procedures and prompts are provided in Appendix A.

**Clinical Rationale Corpus Curation** Our process begins with a raw dataset  $\mathcal{D}_{\text{raw}}$  aggregated from MedQA (Jin et al., 2021), MedCase (Wu et al., 2025b), and JAMA Challenge sources. To isolate the challenging cases that are most valuable for training, we employ an ensemble-based filtering method (Ankner et al., 2025; Mizrahi et al., 2025) where *Llama-3.1-8B-Instruct* (Dubey et al., 2024) attempts each question 16 times, retaining only those answered correctly in fewer than half of the attempts. We then construct high-quality rationales through two complementary pathways. MedCase and JAMA Challenge samples originate from published clinical case reports, providing a rich source of expert-authored reasoning grounded in real patient encounters. For these literature-grounded instances, we use *Gemini-2.5-Pro* (Comanici et al., 2025) to extract and synthesize coherent rationales  $r$  from the original publications. For exam-based instances from MedQA, where no accompanying literature is available, we use *Gemini-2.5-Pro* to generate rationales through iterative sampling with up to 4 attempts, accepting only those that yield the correct answer. This dual-pathway process yields our curated corpus  $\mathcal{D}_{\text{curated}} = \{(q_i, r_i, a_i)\}_{i=1}^N$ .

**Evidence Graph Construction** Next, we transform each textual rationale  $r$  from  $\mathcal{D}_{\text{curated}}$  into a structured Evidence Graph  $G$  via multi-model consensus. Specifically, we employ an ensemble of three diverse models from independent providers (i.e., *GPT-OSS-120B* (Agarwal et al., 2025), *Qwen3-235B* (Yang et al., 2025), and *DeepSeek-V3* (Liu et al., 2024)) to independently extract candidate reasoning triplets from each rationale using a specialized prompt that guides the extraction of subject-predicate-object triplets  $t = (s, p, o)$  capturing logical connections between clinical concepts. To ensure reliability and suppress model-specific artifacts, a triplet is accepted only if it is independently extracted by at least two of the three models, effectively filtering out idiosyncratic noise while retaining consensus-validated reasoning structures. This multi-model consensus process yields our dataset with EGs,  $\mathcal{D}_{\text{EG}} = \{(q_i, r_i, a_i, G_i)\}_{i=1}^N$ , where  $G_i = \{(s, p, o)\}$  is the set of consensus-accepted reasoning triplets for  $r_i$ . In essence, this step robustly converts free-form clinical text into a structured, consensus-grounded representation of its underlying diagnostic logic.

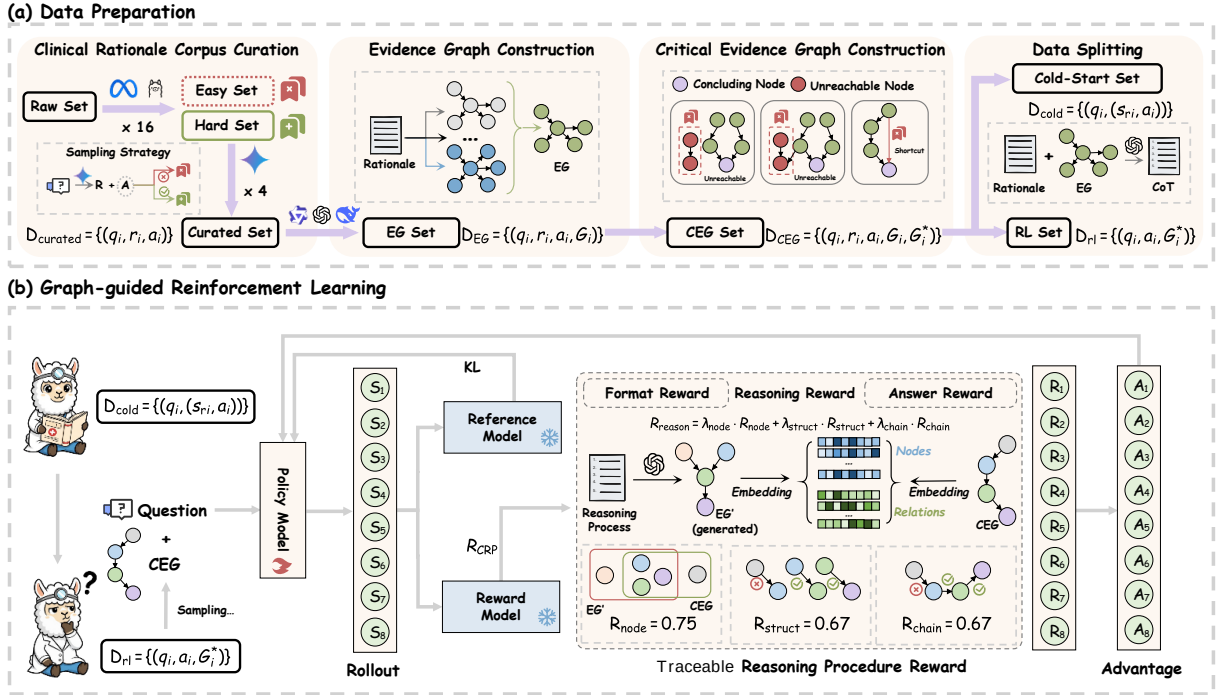


Figure 2: Overview of the MedCEG pipeline. (a) Dataset preparation. Illustrates the process of curating a clinical rationale corpus and subsequently constructing the critical evidence graph. (b) Graph-guided RL. A policy model generates reasoning paths and is optimized using a TRP Reward guided by the critical evidence graph.

**Critical Evidence Graph Extraction** To distill the essential reasoning pathway from the broader EG, we introduce the Critical Evidence Graph  $G^*$ , a refined subgraph that retains only the most granular and logically necessary components for reaching the correct conclusion. Given an Evidence Graph  $G$ , the construction of  $G^*$  proceeds in two stages. First, we identify the conclusion node with maximal semantic similarity to the ground-truth answer, from which we perform a backward traversal over  $G$  to extract the entire causally connected subgraph leading to this conclusion. Second, we apply transitive reduction to this subgraph, which eliminates inferential shortcuts by removing redundant edges, thereby preserving the complete chain of intermediate reasoning steps and preventing any logical leap from being overlooked. We then construct the final dataset  $\mathcal{D}_{\text{CEG}} = \{(q_i, r_i, a_i, G_i, G_i^*)\}_{i=1}^N$ .

**Data Splitting** To facilitate the two-stage progressive learning paradigm (Guo et al., 2025), our final dataset is partitioned into a cold-start subset  $\mathcal{D}_{\text{cold}}$ , which contains CEG-rewritten chain-of-thought sequences for supervised fine-tuning, and an RL subset  $\mathcal{D}_{\text{rl}}$ , split via stratified random sampling based on CEG hops to maintain a balanced distribution of reasoning complexity.

## 4 Graph-Guided Reinforcement Learning

To ensure the logical soundness of our model’s reasoning, we employ a two-stage progressive learning paradigm. The first stage establishes foundational capabilities through a cold-start stage on  $\mathcal{D}_{\text{cold}}$ , a standard methodology detailed in Appendix C. Building on this foundation, we introduce CEG-guided RL that moves beyond outcome-only supervision. Our approach directly addresses this by leveraging the CEG  $G^*$  as a structured reference to evaluate not just what the model concludes, but how it reasons through the clinical evidence.

### 4.1 Traceable Reasoning Procedure Reward

A clinically sound reasoning process must identify the relevant clinical concepts, establish accurate relationships among them, and weave these relationships into a coherent diagnostic narrative. We design a composite Traceable Reasoning Procedure (TRP) reward that mirrors this clinical desideratum through three deterministic graph operations, each computed entirely via rule-based algorithms.

**Semantic Matching ( $R_{\text{node}}$ )** A clinician’s reasoning must be grounded in the essential clinical concepts of a case, from presenting symptoms and laboratory findings to candidate diagnoses. The

semantic matching score quantifies how comprehensively the model’s reasoning covers the critical concepts in the CEG by measuring the average maximum cosine similarity between ground-truth entities  $\mathcal{V}^*$  and generated entities  $\tilde{\mathcal{V}}$ :

$$R_{\text{node}} = \frac{1}{|\mathcal{V}^*|} \sum_{u \in \mathcal{V}^*} \max_{v \in \tilde{\mathcal{V}}} \{\text{sim}(u, v)\}, \quad (1)$$

where  $\text{sim}(u, v)$  is defined as the cosine similarity between the vector representations of the clinical concepts  $u$  and  $v$ . A high  $R_{\text{node}}$  indicates that the model has identified a comprehensive evidential foundation before reaching its conclusion.

**Triplet Recall ( $R_{\text{structure}}$ )** Identifying the right concepts is necessary but not sufficient. The triplet recall reward evaluates whether the model reproduces the factual relationships encoded in the CEG by computing the recall of ground-truth triplets:

$$R_{\text{structure}} = \frac{1}{|G^*|} \sum_{t^* \in G^*} \mathbb{I}(t^* \in \tilde{G}), \quad (2)$$

where  $\mathbb{I}(\cdot)$  is an indicator function that returns 1 if triplet  $t^* = (s, p, o)$  from  $G^*$  is matched in the generated graph  $\tilde{G}$  via semantic similarity thresholding on both entities and relations. This metric penalizes reasoning that links correct concepts through factually incorrect pathways.

**Connected-Component Analysis ( $R_{\text{chain}}$ )** Clinical diagnosis requires not merely a collection of isolated observations but a continuous chain of inference from initial presentation to final conclusion. We assess this logical continuity by constructing an undirected graph from the recalled triplets and measuring the fraction of ground-truth triplets contained within the largest connected component  $C_{\text{max}}$ :

$$R_{\text{chain}} = \frac{|C_{\text{max}} \cap G^*|}{|G^*|}. \quad (3)$$

A high  $R_{\text{chain}}$  confirms that the model has woven its clinical observations into a coherent narrative rather than producing fragmented reasoning islands.

**Final Reward Calculation** The three graph-based metrics are combined into the process reward  $R_{\text{reason}}$ , providing a holistic assessment of reasoning quality that spans conceptual coverage, relational accuracy, and logical coherence:

$$R_{\text{reason}} = \lambda_{\text{node}} \cdot R_{\text{node}} + \lambda_{\text{struct.}} \cdot R_{\text{struct.}} + \lambda_{\text{chain}} \cdot R_{\text{chain}}, \quad (4)$$

where  $\lambda_{\text{node}}$ ,  $\lambda_{\text{structure}}$ , and  $\lambda_{\text{chain}}$  are weighting coefficients that balance the contribution of each component. This process reward is then integrated with binary rewards for final answer accuracy and output format adherence forming the TRP signal.

## 4.2 Optimization

We employ the group relative policy optimization (GRPO) algorithm to fine-tune the model using the composite TRP reward signal. GRPO iteratively improves the policy  $\pi_{\theta}$  by maximizing expected rewards while penalizing deviations from a reference policy  $\pi_{\text{ref}}$ . For each question  $q$ , the old policy  $\pi_{\theta_{\text{old}}}$  generates a group of  $G$  candidate outputs  $\{o_i\}_{i=1}^G$ . For the  $t$ -th token of the  $i$ -th output, we compute the probability ratio  $r_t(\theta) = \pi_{\theta}(o_{i,t} | q, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})$  and then apply a clipped surrogate objective  $\mathcal{L}_{i,t}^{\text{clip}}(\theta) = \min \left[ r_t(\theta) \hat{A}_{i,t}, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) \hat{A}_{i,t} \right]$ , where  $\hat{A}_{i,t}$  is the advantage estimated within the group and  $\varepsilon$  is the clipping parameter that ensures training stability. The overall GRPO objective is then formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}_{\text{cl}}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left( \mathcal{L}_{i,t}^{\text{clip}}(\theta) - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta} \| \pi_{\text{ref}}] \right) \right], \quad (5)$$

where  $\beta$  is the penalty coefficient for the KL divergence term that constrains the policy update to prevent catastrophic deviation. By optimizing this objective with the TRP reward, GRPO aligns the model’s reasoning process with the structured clinical pathways encoded in the CEG, ensuring that the generated reasoning chains are not only diagnostically accurate but also clinically traceable.

## 5 Experiments

### 5.1 Experimental Setup

**Benchmarks** We perform evaluations on a suite of established medical QA benchmarks organized into two domains. The in-domain benchmarks consist of (i) **MedQA** (Jin et al., 2021), (ii) **MedBullets-5op** (Chen et al., 2025), and (iii) **MedCase** (Wu et al., 2025b). The out-of-domain benchmarks include (iv) **MMLU-H** (Hendrycks et al., 2020), (v) **MMLU-Pro-H** (Wang et al., 2024), and (vi) **DiagArena** (Zhu et al., 2026). Among these, MedCase adopts an open-ended question format,

Table 1: A performance comparison of various medical language models across a diverse suite of benchmarks. Models marked with “†” are based on Llama-3.1-8B-Instruct. The green values in parentheses denote the **absolute performance gain** over the base model. **Bold** denotes the best result, and an underline indicates the second-best.

| Model                           | In-Distribution |                |              |                      | Out-of-Distribution |              |              |                      |
|---------------------------------|-----------------|----------------|--------------|----------------------|---------------------|--------------|--------------|----------------------|
|                                 | MedQA           | MedBullets-5op | MedCase      | Average              | MMLU-H              | MMLU-Pro-H   | DiagArena    | Average              |
| <i>General Models</i>           |                 |                |              |                      |                     |              |              |                      |
| Gemma3-4B-it                    | 38.88           | 37.01          | 11.82        | 29.24                | 44.63               | 24.94        | 21.53        | 30.37                |
| Qwen2.5-7B-Instruct             | 54.28           | 37.99          | 15.16        | 35.81                | 69.24               | 49.39        | 21.96        | 46.86                |
| Llama-3.1-8B-Instruct           | 61.59           | 44.48          | 16.83        | 40.97                | 73.00               | 51.71        | 35.63        | 53.45                |
| <i>Medical Instruct Models</i>  |                 |                |              |                      |                     |              |              |                      |
| MedGemma-4B-it                  | 63.16           | 45.45          | 14.05        | 40.89                | 66.30               | 43.28        | 40.44        | 50.01                |
| OpenBioLLM-8B                   | 43.75           | 32.79          | 12.37        | 29.64                | 54.64               | 28.61        | 39.34        | 40.86                |
| UltraMedical3.0-8B              | 61.57           | 41.56          | 14.94        | 39.36                | 66.67               | 53.30        | 34.10        | 51.36                |
| UltraMedical3.1-8B†             | 73.21           | 54.55          | 15.38        | 47.71 (+6.7)         | 77.04               | 57.82        | 38.69        | 57.85 (+4.4)         |
| <i>Medical Reasoning Models</i> |                 |                |              |                      |                     |              |              |                      |
| MedS <sup>3</sup> -8B†          | 71.88           | 49.68          | 18.84        | 46.80 (+5.8)         | 79.50               | 57.21        | 36.50        | 57.74 (+4.3)         |
| MedReason-8B†                   | 71.80           | 55.50          | 19.06        | 48.79 (+7.8)         | 75.76               | 57.09        | 36.83        | 56.56 (+3.1)         |
| AlphaMed-8B†                    | <b>75.41</b>    | <u>61.69</u>   | –            | 45.70 (+4.7)         | <u>79.87</u>        | <b>65.28</b> | 37.16        | 60.77 (+7.3)         |
| Huatuo-o1-8B†                   | 72.60           | 53.90          | 18.39        | 48.30 (+7.3)         | 79.34               | 58.70        | <u>49.18</u> | 62.41 (+9.0)         |
| <i>Our Models</i>               |                 |                |              |                      |                     |              |              |                      |
| Raw CoT Distillation†           | 70.83           | 56.17          | 22.41        | 49.80 (+8.8)         | 77.34               | 55.29        | 42.17        | 58.27 (+4.8)         |
| MedCEG (only SFT)†              | 73.06           | 58.77          | 28.09        | 53.31 (+12.3)        | 79.31               | 59.54        | 48.63        | <u>62.49</u> (+9.0)  |
| MedCEG (cold start)†            | 71.96           | 59.42          | 28.87        | <u>53.42</u> (+12.5) | 80.47               | 57.82        | 47.54        | 61.94 (+8.5)         |
| MedCEG†                         | <b>75.41</b>    | <b>63.96</b>   | <b>31.88</b> | <b>57.08</b> (+16.1) | <b>81.21</b>        | <u>62.22</u> | <b>50.16</b> | <b>64.53</b> (+11.1) |

while the remaining benchmarks employ multiple-choice question format.

**Baselines** We benchmark MedCEG against three categories of models. (i) **General models** include *Gemma3-4B-it* (Team et al., 2025), *Llama3.1-8B-Instruct* (Dubey et al., 2024), and *Qwen2.5-7B-Instruct* (Qwen et al., 2025). (ii) **Medical instruct models** include *MedGemma-4B-it* (Sjellergren et al., 2025), *OpenBioLLM-8B* (Pal and Sankarasubbu, 2024), *UltraMedical3.0-8B* (Zhang et al., 2024), and *UltraMedical3.1-8B* (Zhang et al., 2024). (iii) **Medical reasoning models** include *MedS<sup>3</sup>-8B* (Jiang et al., 2026) and *Huatuo-o1-8B* (Chen et al., 2024), which are trained with a PRM, *MedReason-8B* (Wu et al., 2025a), which is fine-tuned on Huatuo-o1-8B using a human-designed CoT approach, and *AlphaMed-8B* (Liu et al., 2025a), which only utilizes an outcome reward.

**Evaluation** We evaluate our approach along three dimensions. First, for outcome accuracy, we measure accuracy on multiple-choice questions and employ LLM as an automatic judge for open-ended answers. Second, for reasoning process quality, a committee of three models (*DeepSeek-R1*, *Qwen3-235B*, *GPT-5-High*) rates each reasoning chain on five criteria, including logical coherence, factual accuracy, evidence faithfulness, interpretability & clarity, and information utilization, with human ex-

pert consistency analysis to validate inter-annotator agreement. Third, for ablation analysis, we examine the individual contribution of each TRP reward component, reward hyperparameter sensitivity and so on. Further ablation studies are in Appendix B.

## 5.2 Benchmarking Results

Table 1 presents benchmarking results across In-Distribution (ID) and Out-of-Distribution (OOD) scenarios. Within the scope of open-source medical reasoning models at table, MedCEG achieves the strongest performance with average scores of 57.08% on ID tasks and 64.53% on OOD tasks.

Compared with the *Llama-3.1-8B-Instruct*, MedCEG yields gains of +16.1 on ID and +11.1 on OOD, substantially exceeding all competing methods, including Huatuo-o1 (+7.3/+9.0) and AlphaMed (+4.7/+7.3). This underscores the critical importance of structured process supervision in medical applications where nuanced understanding of clinical contexts and terminology is paramount.

Notably, MedCEG outperforms the competitive *Huatuo-o1-8B* by 8.78 on ID tasks and 2.12 on OOD tasks. The advantage becomes more pronounced on tasks requiring deeper reasoning. On MedBullets-5op, MedCEG achieves 63.96% compared to *AlphaMed-8B*’s 61.69%. On the challenging open-ended MedCase benchmark, MedCEG reaches 31.88%, substantially exceeding all competing models, while *AlphaMed-8B* fails entirely

to generate valid outputs confirming that process supervision is critically important.

A natural question is **whether the observed gains stem primarily from dataset quality rather than the CEG-based framework itself**. To isolate this factor, we include a raw CoT distillation baseline in Table 1 that fine-tunes Llama-3.1-8B-Instruct directly on raw rationales without graph construction or CEG-guided rewriting. This baseline achieves an ID average of 49.80 and an OOD average of 58.27, substantially underperforming MedCEG (SFT only) (53.31 and 62.49) despite using the same underlying rationale source, confirming that the gains are attributable to structured supervision rather than data volume alone. The further improvement from MedCEG over MedCEG (SFT only) additionally validates the marginal contribution of CEG-guided reinforcement learning.

### 5.3 Reasoning Process Assessment

Beyond outcome accuracy, we assess the quality of the reasoning process. As illustrated in Figure 3, for a fine-grained assessment, a cohort of 2,000 correctly adjudicated instances was randomly sampled from each model under review and scored by the evaluator committee across five dimensions. The scoring criteria are described in Appendix D, and concrete examples are presented in Appendix E.

MedCEG demonstrates clear superiority across all evaluated dimensions, achieving an aggregate score of 8.64 and outperforming the strongest baseline *MedReason-8B* (7.89) by 9.5%. Models such as *MedReason-8B*, which was fine-tuned on meticulously curated datasets, and those employing PRM (*Huatuo-o1-8B* and *MedS<sup>3</sup>-8B*), deliver competitive performance but do not attain the holistic excellence demonstrated by MedCEG. A particularly notable observation concerns *AlphaMed-8B*, which, despite its commendable accuracy on multiple-choice benchmarks, scores as low as 0.98 in Logical Coherence. This gap starkly illustrates the inherent limitations of supervision paradigms that optimize exclusively for final outcomes while neglecting the fidelity of reasoning pathways.

#### 5.3.1 Human Expert Validation

To validate the automated scoring pipeline, we conduct a human expert study involving five medical professionals from the same academic institution. A total of 1,000 instances (200 per model  $\times$  5 models) were randomly shuffled, blinded to model identity, and independently scored across five di-

mensions. To quantify annotation consistency, 100 overlapping instances were annotated by all experts, yielding an average Fleiss’  $\kappa$  of 0.62.

Table 2: Human evaluation results (1,000 instances, 5,000 total judgments). Average reliability within  $\pm 0.5$  reaches 73.6%, with Spearman  $\rho = 0.81$  ( $p < 0.001$ ) between human and LLM committee scores.

| Dimension                  | LLM (Avg.) | Human (Avg.) | Reliability ( $\pm 0.5$ ) |
|----------------------------|------------|--------------|---------------------------|
| Logical Coherence          | 1.74       | 1.71         | 74.8%                     |
| Factual Accuracy           | 1.68       | 1.65         | 73.2%                     |
| Evidence Faithfulness      | 1.71       | 1.69         | 75.4%                     |
| Interpretability & Clarity | 1.65       | 1.63         | 71.0%                     |
| Information Utilization    | 1.62       | 1.60         | 73.6%                     |
| Average                    | 1.68       | 1.66         | 73.6%                     |

As shown in Table 2 and Figure 4, the average reliability reaches 73.6% under a  $\pm 0.5$  threshold, with median score differences close to zero and no systematic directional bias. Analysis of the 264 discrepant instances reveals complementary error patterns rather than systematic bias, with LLMs tending to over-score Factual Accuracy on rare diseases (43%) while under-scoring Interpretability & Clarity (31%). These findings confirm that automated committee scoring serves as a reliable proxy for expert human judgment.

### 5.4 Ablation Studies and Sensitivity Analysis

**Ablation on the Training Pipeline** Table 1 illustrates the impact of each training phase. The SFT-only baseline already surpasses most leading models (53.31 ID, 62.49 OOD), and CEG-guided RL further elevates performance to 57.08 and 64.53, confirming the marginal contribution of process-level rewards. Comparing MedCEG (SFT only) against the Raw CoT Distillation baseline, which uses the same frontier model rationales but without graph construction, reveals a consistent advantage of +3.51 on ID and +4.22 on OOD, directly demonstrating that structured supervision contributes substantially beyond data quality.

**Ablation on the Reasoning Reward** As shown in Figure 5, removing any individual reward component degrades performance, with  $R_{\text{chain}}$  causing the largest drop (ID  $\Delta = 0.46$ , OOD  $\Delta = 1.75$ ), highlighting the importance of step-by-step logical coherence. Relying solely on outcome supervision (w/o  $R_{\text{reason}}$ ) yields only marginal gains over the initial stage. Notably, replacing CEG-based rewards with the denser EG-based framework (w/  $\mathcal{D}_{\text{EGs}}$ ) causes performance to collapse (ID 55.18, OOD 61.23) despite using identical data.

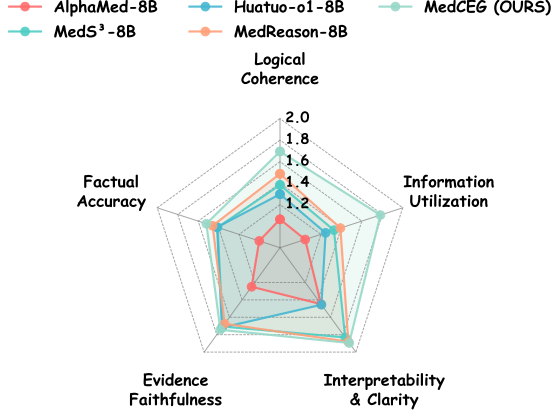


Figure 3: Multi-dimensional evaluation of the reasoning process for different models. This chart compares the quality of the reasoning process for MedCEG against four baseline models across five clinically relevant dimensions.

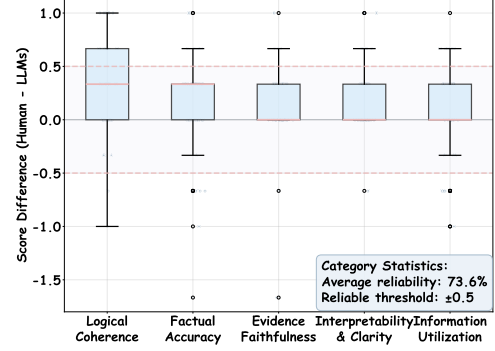


Figure 4: Box plot analysis of scoring consistency between human and LLM evaluations. The y-axis represents the score difference, where positive values indicate higher scores from humans while negatives indicate lower scores.

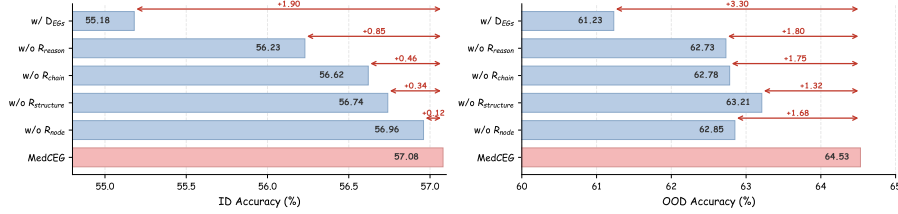


Figure 5: Ablation study on the reasoning reward. We compare our full model against several variants.  $w/ \mathcal{D}_{EGs}$  uses EG data for RL instead of CEG data,  $w/o R_{reason}$  removes the entire reasoning reward,  $w/o R_{chain}$ ,  $w/o R_{structure}$ , and  $w/o R_{node}$  ablate the reward modules for semantic matching, triplet recall and connected-component analysis.

Table 3: Sensitivity analysis on semantic matching.

|          | Varying $\theta_{ent}$ (fixed $\theta_{rel} = 0.70$ ) |              |       | Varying $\theta_{rel}$ (fixed $\theta_{ent} = 0.80$ ) |              |       |
|----------|-------------------------------------------------------|--------------|-------|-------------------------------------------------------|--------------|-------|
|          | 0.70                                                  | 0.80         | 0.90  | 0.60                                                  | 0.70         | 0.80  |
| ID Acc.  | 56.89                                                 | <b>57.34</b> | 56.97 | 56.94                                                 | <b>57.34</b> | 56.72 |
| OOD Acc. | 62.67                                                 | <b>63.12</b> | 62.74 | 62.81                                                 | <b>63.12</b> | 62.49 |

**Sensitivity of Semantic Matching** The semantic threshold controls whether a generated node or triplet is considered a valid match to a ground-truth CEG element. We vary  $\theta_{entity}$  (fixing  $\theta_{relation} = 0.70$ ) and  $\theta_{relation}$  (fixing  $\theta_{entity} = 0.80$ ) separately, each trained to 60 RL steps. As shown in Table 3, performance remains stable across a moderate threshold range, peaking at our defaults ( $\theta_{entity} = 0.80$ ,  $\theta_{relation} = 0.70$ ). Overly permissive thresholds introduce spurious matches that corrupt the reward signal, whereas overly stringent ones penalize valid but paraphrased outputs.

**Sensitivity of Reward Component** The TRP reward combines  $R_{node}$ ,  $R_{struct}$ , and  $R_{chain}$  with weights  $\lambda_{node}$ ,  $\lambda_{struct}$ , and  $\lambda_{chain}$ . We conduct a grid search over representative weight combina-

Table 4: Sensitivity analysis on reward component weights. The composite reward is robust to moderate perturbations in weight selection.

|          | Weight combinations ( $\lambda_{node}, \lambda_{struct}, \lambda_{chain}$ ) |                    |                           |                    |
|----------|-----------------------------------------------------------------------------|--------------------|---------------------------|--------------------|
|          | (0.40, 0.20, 0.40)                                                          | (0.45, 0.25, 0.30) | <b>(0.50, 0.30, 0.20)</b> | (0.55, 0.35, 0.10) |
| ID Acc.  | 56.96                                                                       | 57.01              | <b>57.34</b>              | 56.83              |
| OOD Acc. | 62.49                                                                       | 62.84              | <b>63.12</b>              | 62.31              |

tions while fixing their sum to 1.0, each trained to 60 RL steps. As shown in Table 4, performance remains consistently strong across a range of weight combinations, suggesting that no single component dominates. The default setting ( $\lambda_{node}=0.50$ ,  $\lambda_{struct}=0.30$ ,  $\lambda_{chain}=0.20$ ) achieves the best overall accuracy, indicating that conceptual coverage is slightly more influential than others in our composite reward. We provide more ablation studies and sensitivity analysis experiments in Appendix B.

## 6 Conclusion

In this paper, we present MedCEG, a graph guided framework for reinforcing clinically traceable medical reasoning by supervising intermediate steps

with Critical Evidence Graphs. We construct CEGs from challenging clinical cases and optimize small, locally deployable models with a process reward that jointly measures semantic matching, triplet recall and connected-component analysis. Experiments across in domain and out of domain benchmarks show consistent accuracy gains and improved reasoning quality, and additional ablations and sensitivity analyses support the robustness of our reward design.

## Limitations

Our study has several limitations. First, while Med-CEG achieves competitive performance among open-source models of comparable scale, a gap remains when compared with substantially larger proprietary models. Further scaling of data and model capacity may help bridge this gap. Second, the construction of Critical Evidence Graphs depends on frontier models for triplet extraction, which may introduce subtle biases despite our multi-model consensus filtering. Third, our evaluations are conducted primarily on English-language medical benchmarks, and the generalizability to other languages remains to be explored.

## Ethical Considerations

**Clinical Safety and Oversight** We design our model strictly as a clinical decision support tool, not an autonomous agent. While our *Critical Evidence Graph* (CEG) enhances logical transparency, it does not eliminate the risk of hallucinations. Consequently, all outputs must be verified by qualified healthcare professionals. We strongly advise against deployment in clinical settings without a robust “human-in-the-loop” framework.

**Data Privacy** The dataset released in this work is derived exclusively from publicly available, de-identified medical benchmarks. No Protected Health Information (PHI) or private patient data was processed, ensuring compliance with data privacy regulations and ethical standards.

## References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max M Marion, Matthew Leavitt, and Mansheej Paul. 2025. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *International Conference on Learning Representations*, volume 2025, pages 100615–100634.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific data*, 10(1):67.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ethan Goh, Robert J Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A Freed, Joséphine A Cool, Zahir Kanjee, Kathleen P Lane, Andrew S Parsons, and 1 others. 2025. Gpt-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine*, 31(4):1233–1238.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

- Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey—part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.
- Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2026. Meds<sup>3</sup>: Towards medical slow thinking with self-evolved soft dual-sided process supervision. 40(37):31319–31327.
- 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.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.
- Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, and 1 others. 2025b. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.
- David Mizrahi, Anders Boesen Lindbo Larsen, Jesse Allardice, Suzie Petryk, Yuri Gorokhov, Jeffrey Li, Alex Fang, Josh Gardner, Tom Gunter, and Afshin Dehghan. 2025. Language models improve when pretraining data matches target tasks. *arXiv preprint arXiv:2507.12466*.
- Malaikannan Sankarasubbu Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-8B>.
- Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 1(2):3.
- 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.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. pages 1279–1297.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.
- Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and 1 others. 2025. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26457–26478.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- 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*, pages 1–9.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025a. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

- Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025b. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.
- Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. Medrlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.
- Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2026. Diagnosisarena: benchmarking diagnostic reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 3074–3098.

## A Details of Dataset Preparation

### A.1 Clinical Rationale Corpus Curation

First, to construct the Hard Dataset, we first implement a filtering protocol on the Raw Dataset  $\mathcal{D}_{\text{raw}} = \{(q_i, a_i)\}$  to isolate the most challenging clinical scenarios. This process involves running inference on each instance using *Llama3.1-8B-Instruct* for  $N = 16$  independent iterations with the prompt detailed in Figure 6. Any instance that is answered correctly in more than half of these trials is considered non-trivial and is subsequently excluded. This selection methodology ensures that the Hard Dataset is populated exclusively with complex cases that demand robust reasoning.

Subsequently, we construct high-quality rationales through two complementary pathways. MedCase and JAMA Challenge samples originate from published clinical case reports that contain expert-authored reasoning grounded in real patient encounters. For these literature-grounded instances, we use *Gemini-2.5-Pro* to extract and synthesize coherent rationales from the original publications. For exam-based instances from MedQA, where no accompanying literature is available, we use *Gemini-2.5-Pro* to perform rejection sampling with up to 4 attempts, retaining only the first response that yields the correct answer along with its corresponding analysis. This dual-pathway process yields our curated corpus  $\mathcal{D}_{\text{curated}} = (q_i, r_i, a_i)$ . The specific prompts used for literature extraction and rejection sampling are provided in Figure 8 and Figure 7, respectively.

### A.2 Details of Evidence Graphs Construction

To construct a comprehensive reasoning graph for each process, we employ the meticulously designed prompt detailed below for extraction. This procedure involves retaining triplets that exhibit high consistency across multiple LLMs, ultimately yielding the dataset  $\mathcal{D}_{\text{EG}} = \{(q_i, r_i, a_i, G_i)\}$ . The specific prompt used for triplet extraction is provided in Figure 9.

### A.3 Details of Thinking Process Rewriting

To ensure the generated rationales maintain natural language flow while preserving the integrity of the underlying inferential logic, we employ *GPT-OSS-120B* to transform each reasoning graph  $G$  into a logically coherent reasoning sequences  $s_r$  in natural language through prompt in Figure 10.

**Question:** {{Question}}

Please reasoning step by step, and put your final answer in \boxed{ }.

Figure 6: Prompt to Construct the Hard Dataset

Provide a comprehensive and well-reasoned answer to the following question. Your analysis must be thorough, leading logically to a definitive conclusion.

**Question:** {{Question}}

**Instructions:**

1. Deconstruct the Question: Begin by breaking down the question into its core components and defining any key terms to establish a clear foundation for your analysis.
2. Develop a Comprehensive rationale:
  - Present a detailed, step-by-step analysis that methodically addresses all parts of the question. Your reasoning must be transparent and easy to follow.
  - For Case-Specific Questions: Dissect the case by identifying the critical facts, underlying principles, and relevant context. Analyze how these elements interact to logically lead to the outcome.
  - Ensure your entire rationale forms a coherent and persuasive argument that directly supports your final conclusion.
3. Formulate the final\_answer:
  - Start with a concise summary of the main points from your rationale.
  - Conclude with a direct and unambiguous answer to the original question.
  - The final, conclusive answer must be enclosed within the \boxed{ }.

**Output Format:**

Your entire response must be a single JSON object with the keys “final\_answer” and “rationale”, strictly adhering to the following structure:

```
{“final_answer”: ..., “rationale”: ...}
```

Figure 7: Prompt to Construct the Curated Dataset

#### A.4 Illustrative Examples of the Training Data

In this subsection, we present a detailed overview of the datasets before and after preprocessing, as summarized in Table 5. After preprocessing, the resulting data are further divided into the supervised fine-tuning set and the reinforcement learning set using an 8:2 ratio.

Table 5: Statistics of the datasets before and after preprocessing.

| Dataset | Original Size | Final Size |
|---------|---------------|------------|
| MedQA   | 13.1k         | 1.1k       |
| MedCase | 10.2k         | 7.3k       |
| JAMA    | 2.1k          | 1.9k       |

## B Additional Ablation and Analysis

This section presents additional ablation studies that complement the core analyses in Section 5.4.

Table 6: Cross-benchmark deduplication analysis. All training-evaluation pairs with potential overlap risk are verified to have zero shared instances after DOI-based filtering.

| Source Pair                          | Dedup Method | Overlap |
|--------------------------------------|--------------|---------|
| JAMA (train) vs. DiagArena (eval)    | DOI matching | 0       |
| MedCase (train) vs. DiagArena (eval) | DOI matching | 0       |
| MedQA (train) vs. MedQA (test)       | Split-level  | 0       |
| MedCase (train) vs. MedCase (test)   | Split-level  | 0       |

We examine the deduplication analysis across data sources, the impact of directed versus undirected connected-component analysis formulations, the generalizability of MedCEG across different model families and parameter scales, and the contribution of multi-LLM consensus filtering compared to single-model distillation.

### B.1 Deduplication Analysis

To ensure a fair evaluation, we conduct a rigorous deduplication analysis across all data sources.

You are a medical expert. Given a clinical question and its source literature, extract and synthesize a coherent rationale from the publication that logically leads to the correct answer.

**Question:** {{Question}}

**Source Literature:** {{Literature}}

**Instructions:**

1. Identify Key Evidence: Carefully read the provided literature and identify the clinical findings, diagnostic results, and reasoning steps that are directly relevant to answering the question.
2. Synthesize a Coherent Rationale:
  - Organize the extracted evidence into a logical, step-by-step reasoning chain that faithfully reflects the analytical process described in the literature.
  - Preserve the original clinical logic and causal relationships presented by the authors without introducing external information or speculation.
  - Ensure the rationale forms a complete and self-contained argument from initial presentation to final conclusion.
3. Formulate the final\_answer:
  - Provide a concise summary of the key reasoning steps from the synthesized rationale.
  - Conclude with a direct and unambiguous answer to the original question.
  - The final, conclusive answer must be enclosed within the \boxed{...}.

**Output Format:**

Your entire response must be a single JSON object with the keys “final\_answer” and “rationale”, strictly adhering to the following structure:  
 {“final\_answer”: ..., “rationale”: ...}

Figure 8: Prompt to Extract Rationales from Source Literature

By construction, training and test splits of the same benchmark are mutually exclusive. For cross-benchmark pairs with potential overlap risk (e.g., JAMA Challenges vs. DiagArena, MedCase vs. DiagArena), we perform DOI-based deduplication for all paper-sourced data. The results confirm zero overlap in all such pairs. A complete provenance summary is provided in Table 6.

## B.2 Directed vs. Undirected Connected-Component Analysis

An important design choice in our framework is the use of undirected connectivity for connected-component analysis. While a directed graph might seem more faithful to the causal nature of clinical reasoning, directional phrasing of extracted triplets is highly sensitive to surface-form variation. For example, the triplets (CT scan, reveals, supraglottic edema) and (supraglottic edema, is found on, CT scan) encode the same clinical fact but traverse the edge in opposite directions. Enforcing directed matching would introduce substantial false negatives, penalizing models that reason correctly but phrase relationships differently from the ground-truth CEG. To empirically validate this design choice, we compare our undirected formu-

Table 7: Comparison between undirected connectivity and directed longest-path formulations for connected-component analysis.

| Connected-Component Metric            | ID Acc.      | OOD Acc.     |
|---------------------------------------|--------------|--------------|
| Directed longest path                 | 55.89        | 61.53        |
| <b>Undirected connectivity (ours)</b> | <b>57.34</b> | <b>63.12</b> |

lation with a directed longest-path alternative.

As shown in Table 7, the directed alternative underperforms across both metrics (−1.45 ID, −1.59 OOD), consistent with our analysis that surface-form direction mismatches introduce reward noise that impairs RL optimization.

## B.3 Generalizability across Parameter Scales

To verify that MedCEG is not specific to a single model family or parameter scale, we apply the full training pipeline to the Qwen3 series models at 4B and 8B scales.

As shown in Table 8, MedCEG consistently enhances reasoning capabilities across all evaluated benchmarks, achieving substantial average improvements of +3.71 points on Qwen3-4B and +5.78 points on Qwen3-8B. Notably, the method proves particularly effective on MedBullets, yield-

Table 8: Performance improvements on Qwen3 series models. MedCEG consistently enhances reasoning capabilities across different parameter scales, demonstrating robust effectiveness and substantial gains.

| Model                         | MedQA                | MedBul.              | MMLU                 | MMLU-P.              | Avg.                 |
|-------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| <i>Qwen3-4B-Instruct-2507</i> |                      |                      |                      |                      |                      |
| Base Model                    | 70.38                | 49.68                | 80.90                | 62.96                | 65.98                |
| + MedCEG                      | <b>72.90</b> (+2.52) | <b>59.09</b> (+9.41) | <b>83.20</b> (+2.30) | <b>63.57</b> (+0.61) | <b>69.69</b> (+3.71) |
| <i>Qwen3-8B</i>               |                      |                      |                      |                      |                      |
| Base Model                    | 76.83                | 56.49                | 84.48                | 65.28                | 70.77                |
| + MedCEG                      | <b>82.01</b> (+5.18) | <b>62.99</b> (+6.50) | <b>88.71</b> (+4.23) | <b>72.49</b> (+7.21) | <b>76.55</b> (+5.78) |

Table 9: Comparison between single-model distillation and the multi-LLM consensus pipeline. Both conditions use rationales from the same underlying source, isolating the contribution of consensus filtering.

| Model                  | MedQA        | MedBul.      | MMLU         | MMLU-P.      | Avg.         |
|------------------------|--------------|--------------|--------------|--------------|--------------|
| Base (Llama-3.1-8B)    | 61.59        | 44.48        | 73.00        | 51.71        | 57.70        |
| + Single-Model Distil. | 70.83        | 56.17        | 77.34        | 55.29        | 64.91        |
| + MedCEG (SFT only)    | <b>73.06</b> | <b>58.77</b> | <b>79.25</b> | <b>59.54</b> | <b>67.66</b> |
| + MedCEG (Full)        | <b>75.41</b> | <b>63.64</b> | <b>81.18</b> | <b>62.22</b> | <b>70.61</b> |

ing impressive gains of +9.41 and +6.50 respectively. This consistency indicates that MedCEG is not bound to specific architectures and scales efficiently with larger backbone models, unlocking greater reasoning potential.

#### B.4 Multi-LLM Consensus vs. Single-Model Distillation

To assess whether the multi-LLM consensus mechanism genuinely improves supervision quality rather than merely amplifying shared biases, we compare against a single-model distillation baseline that uses CoT data generated solely by Gemini-2.5-Pro without consensus filtering or graph-based rewriting.

As shown in Table 9, MedCEG (SFT only) consistently outperforms single-model distillation across all benchmarks (+2.75 average), demonstrating that multi-LLM consensus filtering actively suppresses model-specific artifacts rather than merely amplifying shared biases. The full MedCEG pipeline with CEG-guided RL further extends this advantage to +5.70 average, confirming that both the consensus-based data construction and the graph-guided reward contribute complementary improvements. We acknowledge that biases pervasive across the entire biomedical literature remain a potential concern, and targeted expert review of CEGs in sensitive subdomains such as rare diseases represents a meaningful future direction.

#### B.5 Discussion on the Role of Frontier Models

MedCEG uses frontier models only in the offline stages of data construction and CEG extraction, which may raise the concern that its gains mainly come from stronger teacher models. We therefore clarify the role of frontier models and the scope of our contribution.

MedCEG is not intended to compete directly with frontier closed-source models. Instead, it is designed to improve small, open-source, and locally deployable medical reasoning models for practical clinical settings. In our framework, frontier models are used only to construct supervision data and extract CEGs, and they are not required at inference.

The results in Table 1 show that the gains cannot be explained by teacher-model distillation alone. In particular, Raw CoT Distillation already provides a strong baseline, but MedCEG consistently achieves better performance, indicating that the improvement does not come solely from exposure to stronger model outputs. Rather, it comes from the structured supervision introduced by the CEG representation and the subsequent process-level optimization.

#### C Training and Implementation Details

All experiments are conducted on 8 NVIDIA H100 GPUs using the LLaMA-Factory (Zheng et al., 2024) and VERL (Sheng et al., 2025). In the Cold Start Stage, we perform full-parameter fine-tuning on *Llama-3.1-8B-Instruct* with a learning rate (lr) of  $1 \times 10^{-6}$ , optimized using DeepSpeed ZeRO-2. The RL Stage is optimized with  $lr = 5 \times 10^{-7}$ . A KL-divergence penalty with a coefficient  $\beta = 0.001$  is used to regularize the policy against the SFT model. Both stages adopt a cosine learning rate decay schedule.

The SFT phase was implemented utilizing the *LLaMA-Factory* (Zheng et al., 2024) framework. We conducted full-parameter fine-tuning on the *Llama-3.1-8B-Instruct*. The optimization was performed for 8 epochs, employing a cosine learning rate decay schedule initialized from a peak learning rate of  $1E-6$  with a warmup proportion of 0.1. An effective batch size of 2 was maintained by configuring a per-device batch size of 1 and accumulating gradients over 2 steps. To enhance computational efficiency, we leveraged bfloat16 mixed-precision training and adopted the DeepSpeed ZeRO-2 strategy for distributed execution. This stage was conducted on 8xH100 GPUs and

took 1 hour. A comprehensive summary of the SFT hyperparameters is provided in Table 10.

Table 10: Hyperparameters for the cold-start stage.

| Parameter                      | Value                    |
|--------------------------------|--------------------------|
| <i>Model Configuration</i>     |                          |
| Base Model                     | Llama-3.1-8B-Instruct    |
| Finetuning Strategy            | Full-parameter           |
| <i>Optimization Parameters</i> |                          |
| Learning Rate                  | 1E-6                     |
| LR Scheduler                   | Cosine Annealing         |
| Warmup Proportion              | 0.1                      |
| Training Epochs                | 8.0                      |
| Per-Device Batch Size          | 1                        |
| Gradient Accumulation Steps    | 2                        |
| Optimizer Precision            | BF16                     |
| Distributed Framework          | DeepSpeed (ZeRO Stage 2) |
| <i>Data Handling</i>           |                          |
| Max Sequence Length            | 10,240                   |

The subsequent RL stage was executed using the *VERL* (Sheng et al., 2025) framework, which facilitates our Graph-based Reward Policy Optimization (GRPO) algorithm. Both the actor and the reference policies were initialized from the converged parameters of the model obtained in the Cold Start stage.

The composite Traceable Reasoning Procedure (TRP) reward signal was constructed with the following weights:  $w_{\text{reason}} = 0.3$ ,  $w_{\text{answer}} = 0.6$ ,  $w_{\text{format}} = 0.1$ . The internal components of the reasoning reward,  $R_{\text{reason}}$ , were weighted as  $\lambda_{\text{node}} = 0.5$ ,  $\lambda_{\text{struct}} = 0.3$ , and  $\lambda_{\text{chain}} = 0.2$ . These hyperparameters were determined empirically through preliminary validation experiments. Vector representations of clinical concepts, necessary for the  $R_{\text{node}}$  calculation, were generated using the *bge-large-en-v1.5* embedding model.

The actor policy was optimized over 10 epochs. The learning rate was set to  $5E-7$  with a warmup proportion of 0.185. To maintain training stability and prevent significant policy deviation from the SFT initialization, we incorporated a KL-divergence penalty with a coefficient  $\beta = 0.001$  against the reference policy. The GRPO update step was configured with a global batch size of 32, a mini-batch size of 16, and a per-GPU micro-batch size of 4. During the experience generation (rollout) phase, 8 responses were sampled for each input prompt at a temperature of 1.0. For evaluation, responses were generated deterministically (temperature=0.0). The distributed training process was conducted on a single compute node provisioned with 8xH100 GPUs for 3 days. Detailed

hyperparameters for the RL stage are enumerated in Table 11.

Table 11: Hyperparameters for the RL stage.

| Parameter                                                                | Value                |
|--------------------------------------------------------------------------|----------------------|
| <i>Model &amp; Algorithm</i>                                             |                      |
| Policy Initialization                                                    | Cold Start Model     |
| RL Algorithm                                                             | GRPO (via VERL)      |
| KL Penalty Coefficient ( $\beta$ )                                       | 0.001                |
| <i>Optimization Parameters</i>                                           |                      |
| Actor Learning Rate                                                      | $5.0 \times 10^{-7}$ |
| LR Warmup Proportion                                                     | 0.185                |
| Total Epochs                                                             | 10                   |
| Global Batch Size                                                        | 32                   |
| PPO Mini-batch Size                                                      | 16                   |
| Per-GPU Micro-batch Size                                                 | 4                    |
| <i>Rollout Configuration</i>                                             |                      |
| Generations per Prompt ( $n$ )                                           | 8                    |
| Rollout Temperature                                                      | 1.0                  |
| Evaluation Temperature                                                   | 0.0                  |
| Max Prompt Length                                                        | 4096                 |
| Max Response Length                                                      | 10240                |
| <i>Reward Function Weights</i>                                           |                      |
| $w_{\text{reason}}, w_{\text{answer}}, w_{\text{format}}$                | 0.3, 0.6, 0.1        |
| $\lambda_{\text{node}}, \lambda_{\text{struct}}, \lambda_{\text{chain}}$ | 0.5, 0.3, 0.2        |

## D Clinical Reasoning Quality Assessment

To establish a standardized and quantifiable methodology for the evaluation of clinical reasoning, we developed a comprehensive assessment framework. This framework is engineered to promote and rigorously appraise structured, transparent, and evidence-based clinical thinking. Its intended applications are manifold, encompassing the formative and summative assessment of medical students and resident physicians, performance validation of diagnostic models, quality assurance for Case-Based Discussions, and the retrospective analysis for continuous improvement of clinical decision-making.

### D.1 Five Assessment Dimensions

We implemented this framework by employing a committee of three distinct large language models—namely, *DeepSeek-R1*, *Qwen3-235B*, and *Gemini-2.5-Pro*—to ensure a robust and multifaceted assessment. This committee evaluates the quality of the reasoning process against five core criteria. This multi-dimensional approach aligns with best practices, such as scoring each dimension independently to mitigate the “halo effect.” The five criteria are as follows:

- **Logical Coherence:** This criterion evaluates the internal consistency and deductive validity of the clinical reasoning process. The assessment focuses on whether the thought process unfolds in a structured and rational manner,

moving from evidence to conclusion without logical fallacies or contradictions. A coherent argument should demonstrate a clear chain of reasoning where each step justifiably follows from the previous one.

- **Factual Accuracy:** This dimension verifies the correctness of the medical knowledge that underpins the reasoning. It ensures that the argument is built upon a foundation of established clinical facts, principles, and data. This includes the accurate application of pathophysiology, epidemiology, diagnostic criteria, and understanding of clinical signs and symptoms.
- **Evidence Faithfulness:** This criterion assesses whether the reasoning is strictly grounded in the information provided in the case. The conclusions and intermediate hypotheses must be directly supported by the available evidence (e.g., patient history, physical examination findings, and diagnostic results) without fabricating or “hallucinating” information. This dimension is crucial for ensuring that the reasoning process does not stray from the specific context of the case by making assumptions or introducing external data that was not provided.
- **Interpretability & Clarity:** This evaluates the transparency and comprehensibility of the articulated reasoning. The goal is to determine if a human clinical expert can easily follow the thought process from beginning to end. High-quality reasoning should be presented in a clear, organized, and unambiguous manner, using precise medical terminology correctly. It should avoid convoluted language or poorly structured arguments that obscure the underlying logic. This criterion measures not just the quality of the thinking itself, but also the quality of its communication, ensuring the rationale is transparent and open to scrutiny.
- **Information Utilization:** This dimension measures the thoroughness with which all relevant information from the case is considered and integrated into the final analysis. It assesses whether the reasoning effectively incorporates both key positive findings that support a diagnosis and pertinent negative findings that help rule out others. A high-scoring eval-

uation would demonstrate that no significant piece of data has been overlooked.

The detailed LLM prompts for the five evaluation dimensions and the annotation guidelines for medical professionals are provided in Figures 11, 12, 13, 14, and 15.

## E Case Study

In this section, we present a comparative analysis of the reasoning processes for three models, namely *AlphaMed-8B*, *Huatuo-o1-8B*, and our own model MedCEG. To ensure a fair comparison, we selected a case (Figure 16) where all models arrived at the correct conclusion. For each model, we display its step-by-step reasoning (Figures 17, 19, 21, and 22), followed by a human expert’s evaluation with scores and detailed justifications (Figures 18, 20, and 23).

You are an expert medical knowledge engineer. Your task is to analyze a medical reasoning process and convert it into a highly structured and refined knowledge graph. You must extract, standardize, and logically connect all key medical concepts to produce a final, clean JSON output.

### Processing Rules:

#### 1. \*\*Entity Construction: Standardization & Purity (CRITICAL)\*\*

1.1. \*\*Extract & Standardize\*\*: Identify all medical entities (diseases, symptoms, tests, drugs, biomarkers). Immediately unify all synonymous, near-synonymous, or abbreviated entities (e.g., “the tumor”, “the patient’s mass”) into the **single most medically precise and complete standard name** found in the source text (e.g., “invasive ductal carcinoma”).

1.2. **Enforce Purity**: Entities **MUST** be core nouns or noun phrases only. All modifiers (adjectives, states, locations) must be handled in one of two ways:

1.2.1. **Splitting (Preferred)**: Decompose a complex entity into multiple, atomic triplets.

- **INCORRECT**: [“patient”, “has symptom”, “hard mass in upper outer quadrant of left breast”]

- **CORRECT**: [[“patient”, “has symptom”, “breast mass”], [“breast mass”, “has texture”, “hard”], [“breast mass”, “has location”, “upper outer quadrant of left breast”]]

1.2.2. **Modifier Integration**: Move the modifier into the relationship phrase.

- **INCORRECT**: [“HER2/neu positivity”, “associated with”, “poor prognosis”]

- **CORRECT**: [“HER2/neu receptor”, “positivity is associated with”, “poor prognosis”]

#### 2. **Relationship Definition**

2.1 Use clear, directional verb phrases (e.g., “causes”, “is treated with”, “is diagnosed by”, “is positive for”).

2.2 Explicitly mark negative relationships (e.g., “rules out”, “is not associated with”).

2.3 Retain hypothetical relationships from the reasoning process.

2.4 Convert time information into medical temporal expressions (e.g., “acute”, “chronic”).

2.5 Quantify probabilistic statements with risk levels (e.g., “high risk of”).

#### 3. **Bridging Inference for Logical Gaps**

3.1. Review the reasoning flow to ensure logical completeness.

3.2. Add missing, but clinically essential, procedural steps that connect key events. A common failure is jumping from a symptom to a diagnosis without stating the intervening test.

- **Example**: If the text implies a diagnosis was made from a mass, you must add bridging triplets like [“patient”, “undergoes”, “biopsy”] and [“biopsy”, “was performed on”, “mass”].

#### 4. **Output Format**

Your output must be a single JSON object with key: “triplets”

### Correct Output Example

```
{ { "triplets": [ [ "patient", "has age", "65 years" ], [ "patient", "has gender", "female" ], [ "patient", "has symptom", "mass" ], [ "mass", "has texture", "hard" ], [ "mass", "is", "palpable" ], [ "mass", "located in", "left breast" ], [ "patient", "undergoes", "biopsy" ], [ "biopsy", "was performed on", "mass" ], [ "biopsy", "confirms diagnosis of", "invasive ductal carcinoma" ], [ "invasive ductal carcinoma", "is positive for", "HER2/neu receptor" ], [ "HER2/neu receptor", "positivity predicts", "a poor prognosis" ], [ "invasive ductal carcinoma", "is treated with", "Trastuzumab" ] ] }
```

### Reasoning process

```
{ { rationale }
```

Figure 9: Prompt for EG Extraction

Based on the original question and the reasoning graph, please rewrite the rationale to generate a standardized thought process.

**Original Question:** { {Question} }

**Rationale:** { {Rationale} }

**Reasoning Graph:** { {Reasoning Graph} }

**Thinking Process Example:** { {Thinking Process Example} }

Please provide the rewritten thinking process directly.

Figure 10: Prompt for Thinking Process Rewriting

**Core Definition:** Assesses whether the chain of reasoning from evidence (case information) to conclusion (diagnosis) is complete, sound, and free of contradictions, and whether the final conclusion follows naturally and necessarily from the reasoning process.

**Score 2 (Excellent):** The reasoning chain is logically seamless, and the conclusion is a direct, necessary result of the evidence. The entire argument is solid and convincing from premise to conclusion, with no logical leaps or internal contradictions.

*Example:* “The patient presents with typical ischemic chest pain (crushing, radiating to the left arm, with diaphoresis). The key evidence is the ST-segment elevation in leads V2-V4, which directly localizes and confirms an acute anterior myocardial infarction. Therefore, the final diagnosis is AMI.”

**Score 1 (Adequate):** The reasoning process generally supports the conclusion, but contains minor logical flaws. This may include insufficient justification for secondary points, minor inferential gaps, or a conclusion that is too broad/narrow to be precisely supported by the evidence.

*Example:* “The patient has chest pain and ECG abnormalities, indicating a cardiac issue. The ECG changes are consistent with cardiac ischemia. Therefore, the conclusion is Acute Coronary Syndrome (ACS).”

**Score 0 (Inadequate):** The reasoning process severely contradicts the conclusion, or the reasoning chain contains fundamental logical fallacies. The final answer appears random or incorrect, being completely disconnected from or contradictory to the analysis.

*Example:* “The ST-segment elevation on the ECG strongly suggests AMI. The absence of pleuritic pain also lowers the likelihood of a pulmonary embolism. Therefore, the final diagnosis is pulmonary embolism.”

Figure 11: Prompt for Logical Coherence Dimension

**Core Definition:** Assesses the accuracy of all medical knowledge cited in the reasoning process, ensuring it aligns with current, evidence-based clinical guidelines, textbooks, and consensus.

**Score 2 (Excellent):** All cited medical facts are accurate and current. This includes disease definitions, pathophysiology, diagnostic criteria, interpretation of findings, and treatment principles, all consistent with authoritative sources.

*Example:* “The ECG shows ST-segment elevation in leads V2-V4, a hallmark feature of an acute anterior wall myocardial infarction, which is typically associated with the occlusion of the Left Anterior Descending (LAD) coronary artery”

**Score 1 (Adequate):** Contains non-critical factual errors that do not alter the main diagnostic pathway. For instance, citing a slightly inaccurate statistic or a minor error in a non-essential value range.

*Example:* “The ECG shows ST-segment elevation in leads V2-V4, which is indicative of an acute inferior wall myocardial infarction.”

**Score 0 (Inadequate):** Contains one or more critical factual errors that fundamentally mislead the reasoning process or could lead to patient harm.

*Example:* “ST-segment elevation is a benign early repolarization pattern, common in healthy individuals, and therefore has no clinical significance in this case.”

Figure 12: Prompt for Factual Accuracy Dimension

**Core Definition:** Assesses whether the reasoning is strictly and exclusively based on the information provided in the case, avoiding any fabrication of data (i.e., “hallucination”).

**Score 2 (Excellent):** Every step of the reasoning is explicitly traceable to specific information within the input case. All arguments are directly cited from or based on the source text, with no extrapolation beyond the given evidence.

*Example:* “Based on the case description of ‘retrosternal crushing pain’, ‘radiating to the left arm’, and ‘accompanied by diaphoresis’, an acute cardiac event is highly suspected.”

**Score 1 (Adequate):** The reasoning is primarily based on case information but includes minor, reasonable clinical assumptions or slight misinterpretations of the evidence. It introduces small, clinically plausible details not explicitly stated in the text.

*Example:* “The patient’s chest pain, likely accompanied by shortness of breath, points towards an acute cardiac event.”

**Score 0 (Inadequate):** The reasoning contains clear “hallucinations”, fabricating key information not present in the case and using it as a central pillar for the argument.

*Example:* “Laboratory results show the patient’s troponin level is elevated at 15 ng/mL, confirming myocardial necrosis.”

Figure 13: Prompt for Evidence Faithfulness Dimension

**Core Definition:** Assesses whether the reasoning is presented in a structured, professional, and concise manner that is easily understood by a clinical peer.

**Score 2 (Excellent):** The presentation is clear, well-structured, and uses professional, precise language. It follows a standard clinical logic flow (e.g., presentation → differentials → analysis → conclusion), allowing a peer to effortlessly follow the complete thought process.

*Example:* “1. Clinical Presentation: The patient’s symptoms (crushing chest pain, radiation, diaphoresis) are highly suggestive of cardiac-origin pain. 2. Key Investigations: The ECG finding of ST-segment elevation is definitive evidence for AMI. 3. Conclusion: Integrating the clinical picture and ECG, the diagnosis is clearly AMI.”

**Score 1 (Adequate):** The core idea is understandable, but the presentation is flawed by redundancy, repetition, disorganized structure, or ambiguous language. It requires extra effort from the reader to parse the logic.

*Example:* “The patient has chest pain, very painful, and the EKG is also not good, it has changes. So we think it’s a heart problem, because the pain and the EKG both point to the heart. So it should be a heart attack.”

**Score 0 (Inadequate):** The presentation is convoluted, lacks a logical structure, and is filled with meaningless jargon or inappropriate terminology, making the core reasoning difficult or impossible to understand.

*Example:* “Vectorial changes of myocardial repolarization confirm the electrophysiological basis of transmural ischemia. Therefore, despite the negative evidence of pleuritic pain, the differential of dissection persists. The etiology is thus attributed to a cardiac source, an MI is considered.”

Figure 14: Prompt for Interpretability & Clarity Dimension

**Core Definition:** Assesses how thoroughly all key clinical information was utilized, especially how both positive findings and important **pertinent negatives** were integrated to form the final judgment.

**Score 2 (Excellent):** Comprehensively considers all diagnostically significant positive and negative findings. It not only identifies evidence supporting the conclusion but also explicitly explains how key negative findings help to rule out other relevant diagnoses.

*Example:* “The diagnosis of AMI is based not only on positive findings like crushing chest pain and ST elevation, but is also supported by pertinent negatives: the patient’s denial of pleuritic pain and absence of leg swelling significantly lower the probability of other fatal causes like pulmonary embolism.”

**Score 1 (Adequate):** Focuses on the most critical clinical information to support the conclusion but overlooks some secondary or diagnostically valuable clues (positive or negative). The analysis is not fully comprehensive.

*Example:* “The patient’s crushing chest pain and ST-segment elevation on the ECG are classic signs of an acute myocardial infarction.”

**Score 0 (Inadequate):** Exhibits clear “cherry-picking” behavior. It selectively focuses on evidence that supports a preconceived conclusion while systematically ignoring critical information or pertinent negatives that contradict it.

*Example:* (Assuming the ECG in this case was normal) “The patient’s chest pain is classic crushing, radiating pain, therefore the diagnosis is acute myocardial infarction.”

Figure 15: Prompt for Information Utilization Dimension

#### Question Presented in the Case Study

##### Case Information:

A man in his 20s presented with thickening and swelling of hands after immersion in water, accompanied by burning pruritus for 3 years. The lesions occurred as symmetrical white and transparent papules, white keratosis, excessive wrinkling on the dorsum of hands after water immersion and worsened in summer, and remitted in winter. He was treated with topical tretinoin ointment intermittently at a local clinic with little improvement. The lesions gradually aggravated, and the affected area dispersed to wrists and elbows during past 1.5 years. He attributed these changes to frequent hand washing during the COVID-19 epidemic. No family history of similar skin problems and no history of hyperhidrosis, cystic fibrosis, atopy, or trauma to the hands.

##### Physical Examination:

After immersing both hands in water for 10 mins, there were symmetrical white papules, white hyperkeratosis spots, and excessive wrinkling on the thenar eminence, wrist flexure of both hands, and dorsum of fingers and hands.

##### Diagnostic Tests:

- **Skin Biopsy:** Histopathologic examination revealed remarkable compact orthohyperkeratosis in the epidermis, dilatation of intraepidermal eccrine ducts, and hyperplasia of the eccrine sweat glands.
- **Images:**
  - *Image Title:* Clinical presentation after water immersion. *Description:* A and B – Symmetrical white papules, white hyperkeratosis spots, and excessive wrinkling on hands after water immersion.
  - *Image Title:* Histopathologic features. *Description:* C and D – Hyperkeratosis, dilated eccrine ducts in the epidermis and dermis, and hyperplasia of the eccrine sweat gland (hematoxylin-eosin staining).

##### Options:

A: Aquagenic syringeal acrokeratoderma (ASA)    B: Aquagenic keratoderma  
C: Aquagenic palmoplantar keratoderma    D: Aquagenic wrinkling of the palms

Figure 16: Question Presented in the Case Study

### AlphaMed-8B's Original Response

Let's analyze the information provided step by step.

1. **Patient Presentation:** The patient is a young man who presents with thickening and swelling of his hands after immersion in water, accompanied by burning pruritus. He has symmetrical white and transparent papules, white keratosis, and excessive wrinkling on the dorsum of his hands.

2. **Symptoms:** The key symptoms include symmetrical white papules, hyperkeratosis, and excessive wrinkling of the hands after water immersion. He experiences burning pruritus, particularly in summer, which improves in winter.

3. **Diagnosis:** The patient's condition is known as aquagenic syringeal acrokeratoderma (ASA). This condition involves symmetrical white papules, hyperkeratosis, and wrinkling of the hands after exposure to water.

4. **Diagnostic Tests:** A skin biopsy revealed compact orthohyperkeratosis, dilatation of intraepidermal eccrine ducts, and hyperplasia of the eccrine sweat glands.

5. **Options Provided:**

- A. Aquagenic syringeal acrokeratoderma (ASA)
- B. Aquagenic keratoderma
- C. Aquagenic palmoplantar keratoderma
- D. Aquagenic wrinkling of the palms

6. **Conditions:**

- **ASA:** This condition involves keratoderma, wrinkling of the hands after water exposure, and eccrine gland abnormalities.
- **Aquagenic Keratoderma:** Characterized by keratoderma after water exposure, but doesn't always involve eccrine gland abnormalities.
- **Aquagenic Palmoplantar Keratoderma:** Involves keratoderma on the palms and soles after water exposure.
- **Aquagenic Wrinkling of the Palms:** Involves wrinkling of the palms after water exposure.

7. **Choice Analysis:**

- ASA involves keratoderma, wrinkling after water exposure, and eccrine gland abnormalities, matching the patient's presentation.
- Aquagenic keratoderma refers to keratoderma after water exposure but doesn't specify eccrine gland abnormalities.
- Aquagenic palmoplantar keratoderma involves keratoderma on the palms and soles, not specified in this case.
- Aquagenic wrinkling of the palms involves wrinkling after water exposure but doesn't include keratoderma.

8. **Conclusion:** Based on the information provided, the patient's condition involves symmetrical white papules, hyperkeratosis, wrinkling of the hands after water immersion, and eccrine gland abnormalities. The most likely diagnosis is ASA.

9. **Final Answer:** The final answer is \boxed{A}.

Figure 17: AlphaMed-8B's Reasoning Process and Final Answer

### Scoring Rationale for AlphaMed-8B

**Evidence Faithfulness (Score: 2):** The model perfectly adheres to the provided case information. The description of the patient's presentation, symptoms, and diagnostic test results are all accurately extracted from the source text without any addition or fabrication of data.

**Logical Coherence (Score: 0):** The reasoning process is critically flawed due to circular logic. The model states the final diagnosis (ASA) in step 3, at the very beginning of its analysis, and then re-states it in the conclusion (step 8). The intermediate steps do not build a logical bridge from evidence to conclusion but instead assume the conclusion from the start.

**Factual Accuracy (Score: 0):** The definitions provided in step 5 are factually incorrect and misleading. Aquagenic Syringeal Acrokeratoderma (ASA) is a specific histopathological **variant** of Aquagenic Keratoderma, distinguished by the eccrine duct changes seen in this case. By presenting them as mutually exclusive conditions, the model creates a false dichotomy that undermines the entire differential diagnosis.

**Interpretability & Clarity (Score: 1):** While the core message is understandable, the structure is disorganized and highly redundant. Step 7 ("Choice Analysis") largely repeats the flawed definitions from step 6 without adding any new analytical insight, making the argument inefficient and difficult for a clinical peer to follow.

**Information Utilization (Score: 1):** The model identifies the key positive findings but fails to use them effectively. It does not explicitly highlight that the histopathological evidence of **dilated eccrine ducts** is the single most important clue that elevates a general diagnosis of aquagenic keratoderma to the specific ASA variant, thus missing the core diagnostic step.

Figure 18: Human Expert Scoring Rationale for AlphaMed-8B

### Huatuo-o1-8B's Original Response

#### Thinking:

Alright, let's think about this. We've got a young man in his 20s who's been dealing with some pretty uncomfortable symptoms on his hands. He's got thickening and swelling, and it's been going on for three years. That's a long time to be dealing with this. The symptoms get worse when he's in water and improve when it's cold, which is interesting. That's a big clue.

Now, let's look at what's happening on his skin. He's got these symmetrical white papules and white patches, and his skin gets all wrinkly after he's been in water. That's not just any kind of skin issue. It's like his skin is reacting to the water in a specific way.

The biopsy results are really telling. There's compact hyperkeratosis in the epidermis, which means there's a thickening of the outer layer of the skin. And there's dilation of the eccrine ducts and hyperplasia of the eccrine sweat glands. That's a lot of changes happening in the skin's sweat glands. This is starting to sound like a specific condition.

Let's think about what this could be. Aquagenic syringeal acrokeratoderma, or ASA, is a condition that fits these symptoms. It's known for causing skin changes after water exposure, and it involves the eccrine glands. That's exactly what we're seeing here.

Now, let's compare this to the other options. Aquagenic keratoderma and aquagenic palmoplantar keratoderma both involve skin thickening after water exposure, but they don't specifically mention the eccrine glands. And aquagenic wrinkling of the palms is more about the skin wrinkling without the other symptoms we're seeing.

So, putting it all together, the symptoms, the biopsy results, and the way the condition worsens with water exposure all point to aquagenic syringeal acrokeratoderma. It's the only one that matches everything we're seeing here. Yeah, that's definitely the one. The answer is A.

**Final Response:** The answer is `\boxed{A}`.

Figure 19: Huatuo-o1-8B's Reasoning Process and Final Answer

### Scoring Rationale for Huatuo-o1-8B

**Evidence Faithfulness (Score: 2):** The model demonstrates excellent faithfulness to the case details. Throughout its "Thinking" process, it accurately recalls the patient's history, clinical signs, and specific biopsy findings, using them as the sole basis for its reasoning without hallucinating any information.

**Logical Coherence (Score: 1):** The reasoning follows a generally logical path from evidence to conclusion. However, the argument contains a minor flaw in its differential analysis. While it correctly identifies that other options "don't specifically mention the eccrine glands," it doesn't fully articulate why the presence of dilated ducts is the definitive link to ASA, leaving a small logical gap in its final conclusion.

**Factual Accuracy (Score: 1):** The response avoids making a critical factual error but lacks precision. By stating that other aquagenic keratodermas "don't specifically mention the eccrine glands," it implies they are entirely separate conditions. The more accurate fact is that ASA is a specific clinicopathologic variant of Aquagenic Keratoderma, defined by those very gland changes. This nuance is missed, constituting a non-critical factual weakness.

**Interpretability & Clarity (Score: 1):** The entire "Thinking" section is written in a highly conversational and verbose style ("That's a long time to be dealing with this," "Yeah, that's definitely the one."). While the reasoning is understandable, this informal tone is inappropriate for professional clinical analysis and requires a peer to sift through colloquialisms to extract the core logic.

**Information Utilization (Score: 1):** The model successfully uses most of the key information. However, it fails to make the most critical connection explicitly: it doesn't link the specific histopathologic finding of "dilatation of intraepidermal eccrine ducts" to the "syringeal" (relating to a tube or duct) component of the correct diagnosis. This omission prevents the analysis from being fully comprehensive.

Figure 20: Human Expert Scoring Rationale for Huatuo-o1-8B

### MedCEG's Original Response (Part 1)

1. **Analyze the user's request:** The user has provided a clinical case study of a man in his 20s with specific skin symptoms related to water immersion. They've included details about the presentation, history, physical exam, and histopathology. The user wants me to identify the most likely diagnosis from a list of four options.
2. **Deconstruct the clinical presentation:**
  - **Patient:** Man in his 20s.
  - **Chief Complaint:** Thickening and swelling of hands after immersion in water, accompanied by burning pruritus.
  - **History:**
    - Duration: 3 years.
    - Triggers: Water immersion.
    - Symptoms: Symmetrical white and transparent papules, white keratosis, excessive wrinkling on the dorsum of hands.
    - Aggravating factor: Summer. Relieving factor: Winter.
    - Previous treatment: Topical tretinoin with little improvement.
    - Progression: Gradually aggravated, spread to wrists and elbows over 1.5 years.
    - Associated factors: Attributed to frequent hand washing during the COVID-19 epidemic. No family history, no hyperhidrosis, cystic fibrosis, atopy, or trauma.
  - **Physical Examination:**
    - After 10 minutes of water immersion: Symmetrical white papules, white hyperkeratosis spots, excessive wrinkling on thenar eminence, wrist flexure, and dorsum of fingers and hands. This is the key clinical finding. The description "white papules" and "white hyperkeratosis" is a crucial clue.
  - **Diagnostic Tests:**
    - **Histopathology:** Remarkable compact orthohyperkeratosis in the epidermis, dilatation of intraepidermal eccrine ducts, and hyperplasia of the eccrine sweat glands. This is a very specific finding.
3. **Evaluate the options against the clinical presentation:**
  - **A. Aquagenic syringeal acrokeratoderma (ASA)**
    - **What is it?** A rare skin disorder characterized by acrokeratoderma (thickening of the outer layer of the skin) that appears after water immersion.
    - **Clinical features:** Symmetrical, white, translucent papules and plaques on the palms and sometimes the dorsum of the hands, appearing within minutes of water immersion. Excessive wrinkling is a hallmark. Pruritus can occur.
    - **Histopathology:** Hyperkeratosis, acanthosis, and dilatation of the eccrine ducts. This matches the provided histopathology.
    - **Fit with the case:** Trigger: Water immersion – Yes. Symptoms: White papules, hyperkeratosis, excessive wrinkling – Yes. Location: Dorsum of hands, wrists – Yes. Histopathology: Dilatation of eccrine ducts, hyperkeratosis – Yes. This seems like a very strong candidate.

Figure 21: MedCEG's Reasoning Process (Part 1)

### MedCEG's Original Response (Part 2)

#### 3. Evaluate the options... (Continued)

- **B. Aquagenic keratoderma**
  - **What is it?** A general term for skin thickening of the palms and soles that occurs after water immersion. It's a broader category.
  - **Fit with the case:** This is a possible, but less specific, diagnosis. The description "aquagenic syringeal acrokeratoderma" is more specific and fits the details better. The "syringeal" part points directly to the eccrine glands, which is confirmed by the histopathology. "Aquagenic keratoderma" is a good descriptor but not the most precise diagnosis given the specific features.
- **C. Aquagenic palmoplantar keratoderma**
  - **What is it?** A group of disorders characterized by thickening of the skin on the palms and soles after water immersion.
  - **Fit with the case:** The location (palms and soles) is a bit of a mismatch. The case describes involvement of the dorsum of hands, wrists, and elbows, not specifically the palms.
- **D. Aquagenic wrinkling of the palms**
  - **What is it?** This is another name for Aquagenic Syringeal Acrokeratoderma (ASA). It describes the primary feature of excessive wrinkling that occurs after water immersion on the palms.
  - **Fit with the case:** This is essentially the same entity as option A. The case describes more than just wrinkling; it describes white papules and hyperkeratosis. Therefore, A is a more complete description of the clinical picture.

#### 4. Synthesize and conclude:

- The patient presents with a classic triad of symptoms: (1) water-induced skin changes on the hands, (2) white papules and hyperkeratosis, and (3) excessive wrinkling.
- The histopathology confirms the involvement of the eccrine sweat glands (dilatation of ducts, hyperplasia of glands), which is a key feature of this condition.
- The name "Aquagenic Syringeal Acrokeratoderma" perfectly encapsulates these features: "aqua-" (water), "-genic" (induced by), "-syringeal" (related to sweat glands), and "-acrokeratoderma" (skin thickening).
- The other options are either too general (B) or don't fit the specific clinical and histological details as well (C, D).

#### 5. Final check:

- The case description is a textbook presentation of Aquagenic Syringeal Acrokeratoderma (ASA).
- The key features are all present: water-induced white papules, hyperkeratosis, and wrinkling, along with the characteristic histopathology of eccrine duct dilatation and hyperplasia.

**Final Response:** Based on the clinical presentation and histopathological findings, the final answer is \boxed{A}.

Figure 22: MedCEG's Reasoning Process (Part 2) and Final Answer

### Scoring Rationale for MedCEG

**Logical Coherence (Score: 2):** The model's reasoning is exceptionally clear and sound. It follows a perfect clinical reasoning pathway: deconstructing the case data, systematically evaluating each option against the evidence, and synthesizing the findings into a logical conclusion. There are no fallacies or gaps in the argument.

**Factual Accuracy (Score: 2):** All medical facts presented are accurate. The model correctly defines Aquagenic Syringeal Acrokeratoderma (ASA) and correctly identifies the relationship between the different diagnostic terms, noting that "Aquagenic keratoderma" is a broader category and "Aquagenic wrinkling of the palms" is a synonym for ASA. This demonstrates a precise understanding of the topic.

**Evidence Faithfulness (Score: 2):** The response is perfectly faithful to the source material. The "Deconstruct the clinical presentation" step is a masterclass in extracting and organizing information from the case without adding or fabricating any details, ensuring the entire analysis is grounded in the provided evidence.

**Interpretability & Clarity (Score: 2):** The response is exemplary in its clarity and structure. The use of a multi-step process, clear headings, and nested lists makes the entire thought process transparent and effortless for a clinical peer to follow. The final response is a concise and professional summary of the detailed analysis.

**Information Utilization (Score: 2):** The model utilizes all key pieces of information to their full diagnostic potential. Crucially, it makes the explicit connection between the histopathology ("dilatation of... eccrine ducts") and the term "syringeal" in the correct diagnosis, demonstrating a deep and comprehensive understanding of the pathophysiology and terminology.

Figure 23: Human Expert Scoring Rationale for MedCEG