MedPRMBench: A Fine-grained Benchmark for Process Reward Models in Medical Reasoning

Anonymous ACL submission

Abstract

Process-Level Reward Models (PRMs) are increasingly used to guide complex reasoning in large language models, but existing PRM benchmarks focus mainly on general domains and do not capture the safety-sensitive, knowledge-intensive, and uncertain nature of medical reasoning. We introduce MedPRMBench, to our knowledge the first medical-domain PRM benchmark for step-level error-detection evaluation with fine-grained clinical error types and severity labels. Built from seven medical QA sources through a Clinical Reasoning Blueprint (CRB)-guided construction pipeline, MedPRMBench covers 14 error types across Simplicity, Soundness, and Sensitivity, with a four-level clinical severity scale. It contains 6,500 evaluation questions, 13,000 reasoning chains, and 113,910 step-level labels, plus 6,879 training questions. A benchmark-specific Qwen3-8B PRM baseline, trained only on the MedPRMBench training split and reported with a pre-specified final-epoch checkpoint, achieves 87.1% PRMScore, illustrating the value of domain-matched, taxonomy-aware supervision on this benchmark distribution. When used with self-consistency, the PRM yields modest but consistent downstream medical QA gains. Evaluations across proprietary frontier models, open-source reasoning models, medical-specialized models, and medical PRMs reveal persistent gaps in medical reasoning error detection and provide diagnostic directions for future PRM development.

1 Introduction

Large language models (LLMs) have made substantial progress in complex reasoning (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025). Process-Level Reward Models (PRMs), which score intermediate reasoning steps rather than only final answers, are central to both reinforcement learning-based alignment (Lightman et al., 2023; Ouyang et al., 2022) and test-time compute scaling (Wang et al., 2023; Snell et al., 2024). Existing PRM benchmarks and training recipes, however, have been developed largely for general reasoning domains such as mathematics (Lightman et al., 2023; Uesato et al., 2022), leaving medical reasoning comparatively under-evaluated.

Medical reasoning poses domain-specific challenges for process verification. It is **safety-sensitive**, because the clinical impact of an error can range from a redundant step to a missed contraindication or allergy check. It is **knowledge-intensive**, requiring evidence from pharmacology, pathology, guidelines, and patient context. It is also **uncertain**, involving differential diagnosis, probabilistic inference, and context-dependent treatment choices. These properties require PRMs not only to mark steps as correct or erroneous, but also to recognize clinically meaningful error types and severity.

Prior medical PRM work, including MedPRM (Yun et al., 2025), MedCEG (Mu et al., 2025), and MedS^3 (Jiang et al., 2025), mainly studies how to train medical reasoners or reward models. General process-evaluation benchmarks such as PRMBench (Song et al., 2025) and ProcessBench (Zheng et al., 2024) provide useful step-level protocols, but their taxonomies are centered on mathematical reasoning and do not cover medical-specific failures such as missed safety checks, insufficient differential diagnosis, contextual inapplicability, or skipped prerequisites. This leaves a gap in evaluating whether PRMs can detect fine-grained medical reasoning errors.

We address this gap with **MedPRMBench**, a medical-domain PRM benchmark for step-level error-detection evaluation with clinical error taxonomy and severity labels (Figure 1). MedPRMBench is built from seven medical QA sources through a three-phase pipeline: data curation, Clinical Reasoning Blueprint (CRB) construction, and blueprint-guided candidate error construction with

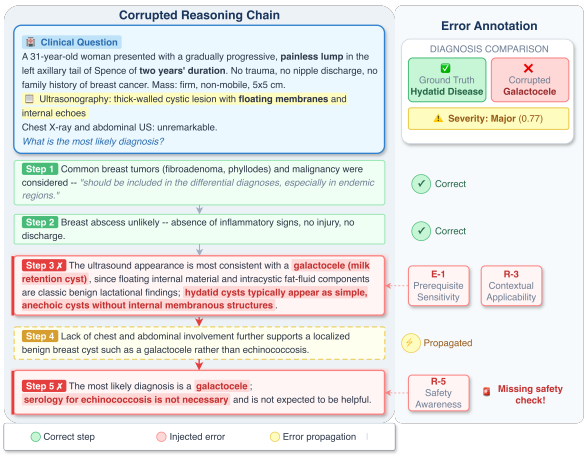

Figure 1: An example from MedPRMBench. The left side shows a medical QA question with its correct reasoning chain (5 steps), while the right side presents the corresponding erroneous reasoning chain where Step 3 and Step 5 are labeled with specific error types (Prerequisite Sensitivity E-1, Contextual Applicability R-3, and Safety Awareness R-5, respectively), along with their severity ratings.

automatic filtering and physician review. The benchmark contains 13,379 questions in total: 6,500 for evaluation, with 13,000 reasoning chains and 113,910 step-level labels, and 6,879 for training. We also train a benchmark-specific Qwen3-8B PRM following the Med-PRM training objective (Yun et al., 2025) and evaluate open-source reasoning models, medical models, proprietary frontier models, and medical PRMs on MedPRMBench.

Our main contributions are as follows:

- We propose **MedPRMBench**, a medical-domain PRM benchmark for step-level reasoning-error detection. The evaluation benchmark comprises 6,500 questions, 13,000 reasoning chains, and 113,910 step-level labels, covering 14 error types and a 4-level severity grading system (Critical, Major, Moderate, Minor), providing a diagnostic resource for medical AI safety evaluation.
- We propose a **blueprint-guided error construction methodology** that uses Clinical Reasoning Blueprints (CRBs) to target clinically meaningful reasoning steps. This approach extracts Evidence Reasoning Networks (ERNs) via multi-model semantic voting and distills them into core causal subgraphs, combined with safety-critical annotation, node criticality computation, and prerequisite annotation to provide structured guidance for candidate error construction and review.
- We design a **medical-domain-specific 14-type error taxonomy** that extends PRMBench's 9 general-purpose categories with medical-specific types including Safety Awareness (R-5), Differential Diagnosis Coverage (E-5), Contextual Applicability (R-3), and Prerequisite Sensitivity (E-1), and uses a 4-level severity grading system to quantify the clinical impact of errors.
- We systematically evaluate open-source reasoning models, medical models, and medical process reward models on MedPRMBench, identifying model-specific weaknesses in medical reasoning error detection and providing diagnostic directions for future PRM improvement.

2 Related Work

Medical process reward modeling has mainly focused on training better medical reasoners or reward models, including retrieval-augmented verification, graph-based clinical evidence alignment, and search-based supervision (Yun et al., 2025; Mu et al., 2025; Jiang et al., 2025). These methods do not directly answer how to evaluate whether PRMs can detect fine-grained medical reasoning errors.

Process-evaluation benchmarks such as ProcessBench (Zheng et al., 2024) and PRMBench (Song et al., 2025) provide step-level or fine-grained evaluation protocols, but they are concentrated in mathematical reasoning. Medical reasoning introduces safety-sensitive consequences, knowledge-intensive evidence integration, uncertainty, and differential-diagnosis requirements that are not captured by existing taxonomies. MedPRMBench fills this evaluation gap with a medical-specific 14-type taxonomy, severity grading, and blueprint-guided error construction. Appendix A provides a detailed related-work discussion and a side-by-side positioning table clarifying that MedPRMBench is primarily an evaluation benchmark, while prior medical PRM works mainly target model training.

3 MedPRMBench Construction

This section presents the construction pipeline of MedPRMBench. As illustrated in Figure 2, the construction process comprises three phases: **Phase 1: Data Curation** (§3.1), **Phase 2: Clinical Reasoning Blueprint Construction** (§3.2), and **Phase 3: Blueprint-Guided Error Construction** (§3.3).

3.1 Data Curation

We construct the source pool by aggregating multiple medical QA benchmarks, applying difficulty-aware filtering, and generating verified step-by-step

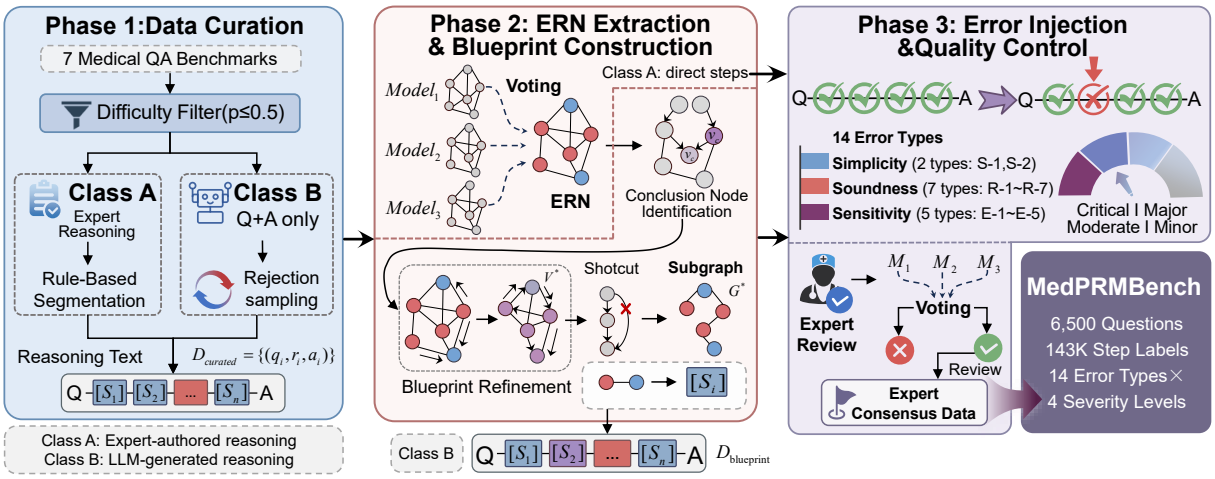

Figure 2: Overview of the MedPRMBench construction pipeline. Phase 1 aggregates data from seven medical QA benchmarks, applies difficulty-aware filtering, and generates reasoning traces via rejection sampling. Phase 2 extracts Evidence Reasoning Networks (ERNs), distills Clinical Reasoning Blueprints (CRBs), and performs safety-critical annotation, node criticality computation, prerequisite annotation, and linearization. Phase 3 leverages the structured blueprint information to construct controlled candidate reasoning errors, followed by quality control and expert revision to produce the final evaluation dataset.

reasoning traces where source explanations are unavailable. The final curated dataset

$$\mathcal{D}_{\text{curated}} = \{(q_i, r_i, a_i)\}_{i=1}^{N} \quad (1)$$

contains challenging questions paired with reasoning supervision for subsequent blueprint construction. Further technical details on blueprint construction and dataset partitioning are provided in Appendix D and Appendix J.

3.2 Clinical Reasoning Blueprints

Free-text reasoning chains, while expressive, obscure the underlying logical structure: which evidence supports which conclusions, and which steps constitute indispensable prerequisites. This makes them difficult to parse and modify in a controlled way. Recent work suggests that explicitly modeling reasoning processes as structured graph representations can support more systematic assessment of reasoning quality (Mu et al., 2025). Inspired by this insight, we construct *Clinical Reasoning Blueprints* (CRBs) as the structural backbone of reasoning chains, enabling subsequent candidate error construction to target clinically relevant causal nodes rather than perturbing random steps.

Specifically, a CRB is a directed knowledge subgraph distilled from reasoning text, representing a core causal reasoning path from clinical evidence to the final conclusion. The CRB, together with its associated safety-critical annotations, prerequisite labels, and node criticality scores, provides structured operational guidance for Phase 3, informing which reasoning steps to target and which error types are clinically applicable. This phase includes reasoning step segmentation, Evidence Reasoning Network extraction, blueprint distillation and verification, safety-critical and prerequisite annotation, and blueprint linearization. An end-to-end CRB construction example is provided in Appendix E.

As described in Section 3.1, datasets are categorized into Class A and Class B based on reasoning text provenance. The two classes follow different processing paths in this phase but converge to a unified output format.

3.2.1 Reasoning Step Segmentation

To enable subsequent processing at a uniform step granularity, we first segment continuous reasoning text into a sequence of independent reasoning steps $\mathbf{s} = (s_1, s_2, \ldots, s_L)$. The two dataset classes employ different segmentation strategies. **Class A**: MedCaseReasoning reasoning text contains explicit numbered formatting (e.g., "1. ... 2. ..."); we apply rule-based regex segmentation. **Class B**: The rejection-sampled reasoning $r^*(x)$ from Phase 1 is structured text; step segmentation is naturally accomplished during subsequent ERN extraction and blueprint linearization.

3.2.2 Evidence Reasoning Network Extraction

For Class B datasets, we extract an *Evidence Reasoning Network* (ERN)—a directed graph $G_{\text{ERN}} = (V, E)$ of medical knowledge triplets from the reasoning text $r^*(x)$, where each edge $e = (v_s, p, v_o) \in E$ connects a subject entity v_s to

an object entity v_o via a clinical relation p (e.g., `suggests`, `treats`, `rules_out`).

To improve extraction quality, we employ a *multi-model semantic voting* strategy: each model in the pool $\mathcal{M} = \{m_1, m_2, m_3\}$ (Qwen3-Max, GPT-5.2, Claude-Opus-4.5) independently extracts triplet sets, and pairwise semantic embedding matching (bge-large-en-v1.5) determines triplet equivalence, retaining only triplets supported by at least $\mu = 2$ distinct models (detailed matching formulas and threshold selection are provided in Appendix D.1). Additionally, the extraction prompt encourages graph connectivity by asking models to link extracted triplets through shared entities when supported by the reasoning text. Class A datasets bypass ERN extraction entirely, using their official reasoning steps directly. This step converts free-text reasoning into a structured representation for subsequent blueprint distillation, yielding the ERN-augmented dataset:

$$\mathcal{D}_{\text{ERN}} = \left\{(q_i, r_i, a_i, G_{\text{ERN},i})\right\}_{i=1}^{N}, \tag{2}$$

where $G_{\text{ERN},i} = (V_i, E_i)$ is the Evidence Reasoning Network constructed from reasoning text r_i.

3.2.3 Blueprint Distillation, Annotation, and Linearization

The ERN typically contains redundant auxiliary information. Blueprint distillation extracts the core causal subgraph from G_{ERN}—the Clinical Reasoning Blueprint $G^* = (V^*, E^*)$—retaining only the critical reasoning path from clinical evidence to the final conclusion. The distillation process comprises four steps:

Conclusion node identification. For a question x with correct answer y^*, we identify the best conclusion node in the ERN via embedding similarity:

$$v_c = \arg\max_{v \in V} \ \text{sim}\big(\phi(v),\ \phi(y^*)\big), \tag{3}$$

where $\phi(\cdot)$ is the bge-large-en-v1.5 encoder. For short answers (e.g., PubMedQA's Yes/No/Maybe), we employ a multi-level fallback strategy: first matching against the explanation text (long_answer), then falling back to the original dataset lookup, and finally to the question text itself.

Semantic bridging pre-connection. Multi-model semantic voting may introduce naming inconsistencies when merging triplets (e.g., different models using different surface forms for the same entity), causing the ERN to fragment into disconnected subgraphs. Before performing BFS, we detect node pairs that refer to the same concept but differ in naming via semantic embedding similarity, and add bridging edges to reconnect fragmented subgraphs, ensuring that subsequent BFS can traverse the complete reasoning path.

Bidirectional BFS subgraph extraction. Starting from the conclusion node v_c, we perform breadth-first search (BFS) on the bridged graph simultaneously along both incoming and outgoing edges, extracting nodes and edges connected to v_c within the reasoning graph:

$$\begin{aligned} V^* = \big\{ v \in V \mid \exists\ \text{path}(v, v_c) \text{ or} \\ \exists\ \text{path}(v_c, v) \text{ in } G_{\text{ERN}} \big\}. \end{aligned} \tag{4}$$

Transitive reduction. We apply transitive reduction (Aho et al., 1972) to the extracted subgraph, removing shortcut edges that can be derived via multi-hop paths, thereby preserving fine-grained reasoning steps.

After distillation, we perform LLM-based *sufficiency verification*: determining whether G^* contains enough information to support the correct answer y^*. For CRBs with insufficient edges (below threshold $\eta_{\min}$), the LLM selects supplementary triplets from the ERN for enhancement, with connectivity filtering requiring newly added triplets to connect to the original CRB. Class A datasets bypass blueprint distillation but undergo the same sufficiency check to assess whether their official reasoning steps support the correct answer. The construction algorithm is provided in Appendix D.3.

Subsequently, we apply *safety-critical annotation* (four levels: Critical/Major/Moderate/Minor), *node criticality computation* (BNC; see Appendix D.4), and *prerequisite annotation*, then linearize the graph structure into a natural-language reasoning step sequence. After the above processing, both classes are unified into:

$$\mathcal{D}_{\text{blueprint}} = \left\{(q_i, \mathbf{s}_i, a_i, G_i^*, \text{SC}_i, \text{Pre}_i)\right\}_{i=1}^{N}, \tag{5}$$

where $\mathbf{s}_i$ is the linearized reasoning step sequence, SC_i denotes safety-critical annotations, and Pre_i denotes prerequisite annotations. This dataset provides the operational foundation for Phase 3 error construction.

3.3 Error Taxonomy and Candidate Construction

Building upon the blueprint dataset constructed in Section 3.2:

$$\mathcal{D}_{\text{blueprint}} = \left\{(q_i, \mathbf{s}_i, a_i, G_i^*, \text{SC}_i, \text{Pre}_i)\right\}_{i=1}^{N}, \quad (6)$$

this phase constructs candidate erroneous reasoning chains from the correct chains and then filters them through verification and expert review. Rather than treating errors as arbitrary text perturbations, the construction is guided by a 14-type error taxonomy spanning three categories—*Simplicity*, *Soundness*, and *Sensitivity*—and proceeds through three core stages: error type selection and target localization, candidate error construction with severity assessment, and composite error synthesis with deterministic verification.

3.3.1 Error Taxonomy

We define 14 error types organized into three categories, each with an inherent severity weight $w_{\text{type}}(c) \in [0, 1]$ (reflecting its potential clinical impact) and a corresponding blueprint operation (Table 1; see Appendix C for detailed illustrations and descriptions of each type).

Table 1: Error taxonomy: 14 types across 3 categories with blueprint operations.

Category	Code	Name	Blueprint Operation	w_{type}
Simplicity	S-1	Non-Redundancy (NR)	Insert redundant step	0.2
	S-2	Non-Circular Logic (NCL)	Inject circular argument	0.3
Soundness	R-1	Evidence-Based Soundness (EBS)	Replace medical fact	0.8
	R-2	Step Consistency (SC)	Introduce contradiction	0.6
	R-3	Contextual Applicability (CA)	Ignore patient context	0.6
	R-4	Confidence Invariance (CI)	Insert overconfident claim	0.7
	R-5	Safety Awareness (SA)	Remove safety check	1.0
	R-6	Information Grounding Compliance (IGC)	Fabricate entity	0.7
	R-7	Trajectory Reasoning (TR)	Reverse causal/temporal order	0.6
Sensitivity	E-1	Prerequisite Sensitivity (PS)	Delete prerequisite step	0.7
	E-2	Deception Resistance (DR)	Insert distractor	0.5
	E-3	Multi-Solution Consistency (MSC)	Dismiss alternatives	0.4
	E-4	Quantitative Correctness (QC)	Alter numerical value	0.5
	E-5	Differential Diagnosis Coverage (DDC)	Narrow differential	0.7

Not all error types are applicable to every question. For each instance, an LLM evaluates the applicability of all 14 types based on the question content, reasoning chain, and blueprint annotations, outputting a binary applicability vector $\mathbf{a}_i \in \{0, 1\}^{14}$. R-1, R-6, and E-2 are designated as universally applicable. From the applicable types, we select 1–3 target error types per instance using distribution-aware sampling, driving the final error type distribution toward the design target π. Detailed sampling formulas are provided in Appendix H.

3.3.2 Candidate Error Construction

All blueprint operations are executed by LLMs using type-specific prompts with few-shot examples, yielding candidate corrupted chains that are subsequently checked by deterministic verification and physician review. To enhance diversity, we employ a *multi-model cross-generation* strategy: a pool of LLMs $\mathcal{M}_{\text{inject}} = \{m_1, m_2, m_3\}$ are used in round-robin fashion. Each LLM receives the original reasoning chain $\mathbf{s}$, the target error type c, the target step index k, and the type-specific prompt, outputting a corrupted reasoning chain $\tilde{\mathbf{s}}$ with metadata indicating modified steps.

For each instance, the operation is executed independently for each selected error type, producing up to 3 *single-error variants*:

$$\mathcal{V}_{\text{single}}(i) = \left\{(\tilde{\mathbf{s}}_i^{(c)}, c, \mathbf{e}_i^{(c)}) \mid c \in \text{selected}(i)\right\}, \quad (7)$$

where $\tilde{\mathbf{s}}_i^{(c)}$ is the corrupted step sequence, c is the error type, and $\mathbf{e}_i^{(c)} \subseteq \{1, \ldots, L\}$ is the set of error step indices.

Each corrupted variant is assigned a four-level severity label (Critical/Major/Moderate/Minor), integrating the disruption fraction $|\mathbf{e}|/L$, the safety criticality weight $w_{\text{SC}}(s_k)$ of the target step, and the inherent severity weight $w_{\text{type}}(c)$ of the error type. For Class B datasets, the BNC value further augments the structural importance measure. The complete scoring formula and discretization rules are provided in Appendix I.

3.3.3 Composite Errors and Verification

To approximate multi-step interacting errors, we synthesize *composite error variants* containing 2–3 single-error types, selecting combinations by cross-category diversity, step disjointness, and severity diversity, with an LLM integrating them into a coherent corrupted chain. We then apply deterministic text diff verification using `SequenceMatcher` to reconcile LLM-reported error indices with actual textual differences, reducing false positives and unreported modifications before physician review. Detailed synthesis strategies and verification algorithms are provided in Appendix F.

After verification and quality control, we retain one corrupted reasoning chain for each evaluation question, paired with its original correct chain. When multiple candidate variants are available, we prioritize variants that satisfy quality-control constraints while improving error-type coverage; multi-error chains are represented with multi-label step-level annotations.

3.4 Quality Control and Statistics

We apply a two-stage quality assurance pipeline combining automatic filtering and full-corpus physician review to improve annotation reliability. The review pool includes approximately 90 physicians with clinical residency training or higher clinical qualifications. Each instance receives one primary physician annotation and one independent physician audit under a blinded protocol, rather than sampled spot-checking; disagreements are adjudicated by a third physician reviewer. The final retained label is physician-adjudicated, with model votes used as an auxiliary flag for additional review and automatic consistency checks rather than as a substitute for expert judgment. Instances with unresolved conflicts or failed consistency checks are excluded from the final dataset.

To make the reliability audit reproducible, the released artifact will include per-instance quality-control metadata: primary label, audit label, adjudicated label, adjudication status, model-vote consistency flag, and filtering reason. These fields support raw agreement, Cohen's κ, Krippendorff's α, adjudication-rate, and filtering-rate analyses for step correctness, error type, and severity labels. We report the construction statistics available in the current manuscript and will include the full audit table with the dataset release.

The resulting MedPRMBench contains 13,379 retained instances with diverse reasoning error types and rich clinical context. Compared to prior PRM benchmarks, it features longer reasoning chains and more complex question formulations, providing a challenging evaluation setting for medical step-level verification. Table 2 summarizes the dataset statistics by error-type occurrence. Because a corrupted reasoning chain may contain multiple injected errors, per-type counts are not mutually exclusive. Detailed quality control procedures, sampling mechanisms, and dataset splitting details are provided in Appendix G, Appendix H, and Appendix J.

4 Evaluation

4.1 Setup

Task. Given a medical question Q and a step-by-step reasoning chain $S = \{s_1, \ldots, s_n\}$, MedPRMBench requires models to assess whether each reasoning step s_i is correct. A step is labeled as erroneous if it contains factual medical mistakes, flawed clinical logic, missing prerequisite reasoning, or contradictions with the clinical vignette. MedPRMBench evaluates every step in the reasoning chain, reflecting the clinical reality that reasoning errors may propagate or co-occur across multiple stages. We evaluate step-level predictions with a balanced PRMScore, the erroneous-step detection rate, and first-error localization accuracy.

Metrics and Evaluation Protocol. For step-level verification, models are asked to assign a binary label to each reasoning step: "+" for a medically correct and logically valid step, and "-" for an erroneous step. For PRM-style models with accessible token logits, we compute the normalized probability over the "+"/"-" tokens at each step boundary and use $P(+)$ as the step-level correctness score. A step is classified as correct if $P(+) \geq 0.5$, and erroneous otherwise. For prompted critic models, including proprietary API models where token-level logits are unavailable, we use the same binary output format and parse the generated "+"/"-" labels as step-level predictions. No checkpoint, prompt variant, or classification threshold is selected using the MedPRMBench test set; all main tables use the basic binary prompt in Appendix N.7. Because probability scores are not available from all API models, threshold-free calibration metrics such as AUROC/AUPRC are not used as the primary comparison, but we release continuous scores wherever available to enable such analyses. **Models.** We evaluate four categories of models: (1) our benchmark-specific medical PRM, (2) the released Med-PRM-Reward v1.0 model, (3) open-source critic models (including both medical and general reasoning models), and (4) proprietary frontier models. Full model descriptions, architectures, and evaluation details are provided in Appendix L.

4.2 Main Results

Evaluation Metrics. Following PRMBench (Song et al., 2025), we adopt **PRMScore** as the primary evaluation metric, defined as:

$$PRM\text{-}Score = w_1 * F1_{neg} + w_2 * F1 \quad (8)$$

where $F1$ treats erroneous steps as the positive class and $F1_{neg}$ treats correct steps as the positive class. We set $w_1 = w_2 = 0.5$ to weight both classes equally, since practical PRMs should detect erroneous steps while avoiding excessive false alarms on correct steps. This metric provides a unified, normalized score that reflects both sensitivity and specificity. We additionally report **Acc**

Table 2: Statistics of MedPRMBench by error-type occurrence. Per-type columns count instances containing each error type and are therefore not mutually exclusive. Abbreviations are defined in Table 1.

	Overall	NR	NCL	EBS	SC	CA	CI	SA	IGC	TR	PS	DR	MSC	QC	DDC
Avg. Steps	9.6	10.9	8.2	9.2	9.5	9.8	9.3	9.7	8.0	9.5	10.0	9.7	12.2	9.4	10.0
Avg. Error Steps	3.3	4.3	3.1	3.5	3.5	3.3	3.4	3.5	3.4	3.6	3.4	3.4	3.2	3.3	3.5
Avg. 1st Error	3.2	3.3	2.8	3.1	3.2	3.2	2.9	3.4	2.7	3.2	3.4	3.2	4.1	3.3	3.0
Avg. Q Length	917	871	892	934	886	996	1076	1040	1018	938	922	823	938	1006	997
# Test Instances	6500	1555	1028	4020	1481	1098	1137	1484	2640	1306	1003	977	668	892	1860
# Instances	13379	3149	1964	8287	2972	2242	2205	2906	5288	2622	2117	2058	1479	1760	3765

(error step detection rate) and **First** (first error step detection accuracy) as supplementary metrics.

The main results are presented in Table 3. We summarize the key findings as follows.

Existing Critic models leave substantial headroom, with differentiated category difficulty. Excluding our benchmark-specific PRM, the average PRMScore across all Critic models is 61.3%. Open-source models mostly hover around 55%, while proprietary frontier models perform better (GPT-5.4 75.4%, Claude-Opus-4.5 73.7%, GPT-5.2 74.6%), yet the results leave substantial headroom. The released Med-PRM-Reward v1.0 baseline reaches 56.1% PRMScore despite high erroneous-step recall (Acc_{err} = 83.8%), indicating that it over-flags many correct steps on MedPRMBench. Across error categories, Critic models perform worst on the Simplicity category (proprietary model average NR 64.5%, NCL 70.4%), suggesting that structural reasoning verification remains difficult for current models.

The benchmark-specific PRM provides an in-distribution upper-bound-style baseline. Our medical PRM achieves an 87.1% overall PRMScore, higher than the strongest Critic model GPT-5.4 by 11.7 pp and the released Med-PRM-Reward v1.0 baseline by 31.0 pp, while also obtaining high Acc_{err} (85.4%) and First (95.5%). This result is obtained with a relatively compact training split of 6,879 questions and 13,758 reasoning chains, whereas Med-PRM-Reward v1.0 uses a larger reward-training corpus of roughly 11.7k questions with multiple sampled traces and retrieved documents (Yun et al., 2025). Because our PRM is trained on the same construction distribution as MedPRMBench, while most critic models are evaluated zero-shot, the comparison should be interpreted as evidence that taxonomy-aware, domain-matched process supervision can fit this benchmark distribution, not as a controlled claim that the model is intrinsically stronger across all natural medical-reasoning errors. As shown in Figure 3, our PRM obtains the highest score on all 14 error types in this evaluation, with the largest margins on the Simplicity category where Critic models are weakest (NR 87.1%, NCL 88.1%). On medical-specific types, our PRM also performs strongly (SA 86.9%, DDC 85.7%, QC 88.0%), supporting the utility of the taxonomy for benchmark-specific training. However, DR (86.0%) remains one of the weakest types, indicating that detecting data dependency errors is still challenging.

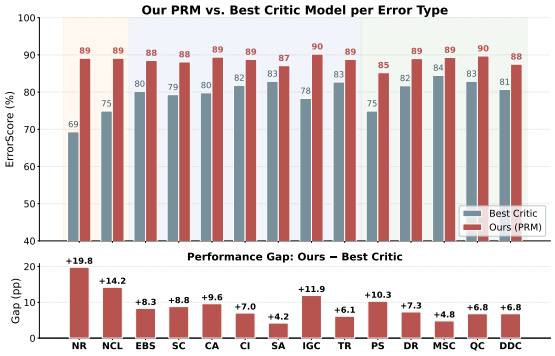

Figure 3: Comparison of our PRM against the best Critic model per error type. The top panel shows absolute PRMScore (%); the bottom panel shows the performance gap (pp). Our PRM obtains higher scores than the best Critic model on all 14 error types, with the largest margins on the Simplicity category (NR, NCL).

4.3 Additional Analyses

We further examine three practical aspects of MedPRMBench evaluation: whether models are biased toward marking steps as correct or erroneous, whether the trained PRM can serve as a plug-and-play verifier for downstream medical QA, and how each construction component contributes to benchmark-specific PRM training. The full tables and discussion are provided in Appendix B. Overall, the analysis shows that frontier models remain unevenly calibrated between correct and erroneous steps, MedPRMBench-RM improves over the average single-trajectory CoT baseline mainly through self-consistency and adds smaller but consistent gains on top of self-consistency, and controlled error construction provides the largest contribution in the ablation study.

Table 3: PRMScore (%) by error type on MedPRMBench. Sub-columns show per-error-type PRMScore across three error categories covering 14 error types; abbreviations are defined in Table 1. The best result in each column is in **bold**; the second best is underlined.

Model	PRMScore	Acc	First	Simplicity		Soundness							Sensitivity				
				NR	NCL	EBS	SC	CA	CI	SA	IGC	TR	PS	DR	MSC	QC	DDC
Open-source Medical Models, Prompted as Critic Models																	
txgemma-9b-chat	57.7	45.5	47.0	55.6	57.9	57.2	58.0	57.8	58.6	58.7	57.1	59.0	58.3	55.6	59.2	58.4	57.9
Llama-3.1-8B-UltraMedical	56.6	43.2	28.7	58.1	55.5	56.1	56.2	55.7	56.5	57.2	56.1	57.1	56.5	55.3	57.6	57.0	56.3
llama-3-meerkat-8b	55.3	37.9	23.6	55.9	54.0	55.0	55.2	54.8	55.1	55.8	54.9	55.9	56.0	54.2	55.9	56.4	54.2
HuatuoGPT-o1-8B	51.8	31.7	65.8	47.4	53.2	51.9	52.7	53.8	54.1	51.1	52.3	50.9	51.0	50.9	49.9	51.5	54.1
Meditron3-8B	46.8	9.7	4.4	43.6	47.0	45.9	46.2	48.4	50.7	46.4	45.9	47.9	48.6	45.3	48.8	48.7	45.7
Avg.	53.6	33.6	33.9	52.1	53.5	53.2	53.7	54.1	55.0	53.8	53.3	54.2	54.1	52.3	54.3	54.4	53.6
Open-source Reasoning Models, Prompted as Critic Models																	
Sky-T1-32B-Flash	65.4	71.2	68.9	61.7	65.8	65.5	65.4	66.7	66.0	67.2	64.1	66.2	66.1	63.1	69.4	66.6	64.9
QwQ-32B	52.4	27.1	54.8	47.7	54.3	52.5	53.3	54.5	53.4	51.6	53.0	52.7	52.5	51.1	50.7	51.1	54.4
Marco-o1	54.3	24.3	25.2	49.3	55.1	53.8	55.7	55.2	57.2	54.1	53.5	55.6	55.9	51.1	56.0	55.8	53.9
DeepSeek-R1-Distill-Llama-70B	51.1	23.3	50.4	46.4	52.6	50.9	51.6	53.9	52.4	50.4	51.5	50.3	51.1	50.0	50.6	51.4	52.4
Qwen3-8B	51.0	19.7	41.4	45.6	53.1	50.9	52.1	53.7	51.8	51.2	51.1	51.5	51.2	50.4	50.5	50.2	52.1
DeepSeek-R1-Distill-Qwen-32B	49.7	17.1	38.0	46.1	50.8	49.4	50.1	52.9	50.0	48.4	50.9	49.0	49.5	49.2	49.5	49.5	51.1
Avg.	54.0	30.4	46.5	49.5	55.3	53.8	54.7	56.1	55.1	53.8	54.0	54.2	54.4	52.5	54.5	54.1	54.8
Proprietary Frontier Models, Prompted as Critic Models																	
Claude-Opus-4.5	73.7	82.5	83.4	69.0	73.6	73.5	73.8	75.2	72.7	76.5	73.9	73.8	74.9	71.4	78.5	76.0	71.2
DeepSeek-V3.2	65.4	81.0	82.6	63.4	65.1	65.1	65.0	66.1	64.1	67.7	64.0	66.1	67.0	62.3	69.0	66.4	64.2
GPT-5.4	75.4	79.0	83.0	68.0	75.4	75.9	76.1	76.8	75.1	78.0	75.2	75.9	75.8	74.0	79.6	77.8	74.0
GPT-5.2	74.6	76.9	81.4	66.3	75.0	75.3	75.5	76.3	75.2	77.3	74.3	74.9	75.1	72.9	79.1	77.0	73.4
GLM-4.7	72.4	76.0	77.3	66.6	72.5	72.6	72.7	73.9	72.7	74.6	72.6	73.5	73.4	69.8	76.3	74.5	70.7
Qwen3-Max	70.4	70.4	74.6	64.3	69.5	70.3	71.1	71.9	71.8	73.6	69.2	71.5	71.9	65.3	76.5	72.4	70.0
DeepSeek-V3.1	70.7	69.0	70.4	64.0	70.5	71.0	71.6	71.9	71.2	74.1	70.1	71.8	72.0	64.7	75.8	74.1	70.0
DeepSeek-R1	67.8	64.0	64.5	63.7	66.8	67.5	67.9	69.6	68.6	70.4	66.1	68.9	68.9	61.6	72.8	70.3	67.4
Gemini-3.1-Pro	63.6	41.5	67.8	55.4	64.8	64.0	64.9	65.6	65.5	63.8	64.6	64.3	63.7	61.3	63.5	65.7	64.8
Avg.	70.4	71.1	76.1	64.5	70.4	70.6	71.0	71.9	70.8	72.9	70.0	71.2	71.4	67.0	74.6	72.7	69.5
Medical Process Reward Models																	
Med-PRM-Reward v1.0	56.1	83.8	65.8	57.1	55.6	58.5	58.5	55.7	56.1	57.1	59.4	58.7	55.7	55.0	53.2	57.9	53.8
Ours	**87.1**	**85.4**	**95.5**	**87.1**	**88.1**	**86.8**	**87.7**	**88.6**	**86.9**	**86.9**	**88.4**	**86.3**	**86.3**	**86.0**	**89.4**	**88.0**	**85.7**

5 Conclusion

In this paper, we investigate whether existing process-level reward models and large language models can detect different types of erroneous reasoning steps in medical contexts. To address this question, we introduce MedPRMBench, a medical-domain PRM benchmark focused on step-level reasoning-error detection. Through a blueprint-guided error construction methodology and a medical-specific 14-type error taxonomy, we curate 13,379 questions (6,500 for evaluation, 6,879 for training) with 13,000 reasoning chains and 113,910 step-level labels, which can be used to evaluate and train process-labeling models. Through comprehensive evaluation of open-source medical models, open-source reasoning models, proprietary frontier models, and medical PRMs, we observe that existing models still leave substantial headroom in medical reasoning error detection, particularly on Simplicity errors and medical-specific error types. The benchmark-specific PRM trained on MedPRMBench provides a strong in-distribution baseline and, when used as an answer-group selector on top of self-consistency, yields modest but consistent downstream QA gains. MedPRMBench is intended to support more standardized process-level reasoning evaluation in medicine and help guide future work on medical PRM assessment and training.

Data and Artifact Release

For the double-blind review version, all author-identifying repository names, acknowledgments, and institutional metadata are removed. The public artifact release will include the train/evaluation split files, source-dataset identifiers and reconstruction scripts where redistribution is license-restricted, the physician annotation guideline, error-injection and evaluation prompts, evaluation scripts, quality-control audit fields, and the benchmark-specific PRM checkpoint. If a source dataset's license does not permit redistributing derived text directly, we will release stable source IDs plus reconstruction code and provide an evaluation server for the corresponding held-out split. The release will preserve the fixed evaluation split used in this paper and will not require using the test set for prompt, threshold, or checkpoint selection.

Limitations

Our work has several limitations. First, MedPRMBench currently focuses on English-language medical reasoning; extending the benchmark to multilingual settings would improve its generalizability across diverse healthcare systems. Second, MedPRMBench is designed as a controlled diagnostic benchmark: it includes model-generated reasoning traces and model-produced candidate errors, but

the final erroneous chains are also structured by CRB-guided operations, deterministic verification, and physician adjudication to improve coverage of specific error types, locations, and severity levels. This design is necessary for fine-grained evaluation, and the errors are therefore not merely template perturbations; nevertheless, the resulting distribution may still differ from fully unconstrained model failures observed in deployment. A complementary held-out subset built from naturally occurring model reasoning errors, with physician-labeled first wrong steps, error types, and severity, would further test out-of-distribution generalization. Third, while our 14-type error taxonomy covers a broad range of medical reasoning defects, it may not exhaustively capture all possible error patterns in specialized clinical domains (e.g., radiology-specific interpretation errors or surgical planning failures). Fourth, the severity grading system, while reviewed by experts and multi-model voting, remains a simplification of the complex, context-dependent nature of clinical error impact. Fifth, our primary metrics use fixed operating points because some API models expose only generated labels rather than calibrated probabilities; threshold-free metrics such as AUROC/AUPRC should be reported for models with continuous scores. Finally, our evaluation is limited to text-based reasoning; extending to multimodal medical reasoning (incorporating medical images, lab reports, etc.) remains an important direction for future work.

Ethics Statement

This research uses publicly available medical QA datasets that do not contain identifiable patient information. All data sources used in MedPRMBench construction are established academic benchmarks with appropriate data use permissions. The expert review process was conducted by qualified medical professionals who provided informed consent for their participation. We acknowledge that medical AI evaluation benchmarks, including MedPRMBench, should not be used as a sole basis for clinical deployment decisions; comprehensive clinical validation remains essential before any AI system is used in patient care settings. Our benchmark is designed to advance research in medical AI safety evaluation and should be used in conjunction with other safety assessment frameworks.

References

Alfred V. Aho, Michael R. Garey, and Jeffrey D. Ullman. 1972. The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137.

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*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yiwen Jiang, Jingyu Chen, Kaijie Zheng, Shunian Huang, Jiageng Zhu, Jie Chen, Shuai Wang, Dahua Zhu, and Zhiqi Chen. 2025. Towards medical slow thinking with self-evolved soft dual-sided process supervision. *arXiv preprint arXiv:2501.12051*.

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.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025a. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

Jiyoun Kim, Hugo J. W. L. Aerts, and Raymond H. Mak. 2025b. Why chain of thought fails in clinical text understanding. *arXiv preprint arXiv:2509.21933*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zicheng Lin, Zhibin Gou, Tian Gong, Zhicheng Liu, Yinghui Wang, Zhengguang Yang, Zhicheng Jiao, Qingjie Cai, Haotian Shi, Yukang Shao, and 1 others. 2024. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 530–546.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024. RM-Bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Yutian Mu, Hao Sun, Jingyi Xu, Jiaqi Gao, Yizhou Ren, Chengqi Zhu, and Jie Zhu. 2025. MedCEG: Reinforcing verifiable medical reasoning with critical evidence graph. *arXiv preprint arXiv:2512.13510*.

NovaSky Team. 2025. Sky-t1: Train your own o1 preview model within $450. Official project page.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 248–260.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, Tahseen Rabbani, Trevor Brokowski, Meditron Medical Doctor Working Group, Tim G. J. Rudner, and Mary-Anne Hartley. 2025. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. In *Proceedings of the AAAI Workshop on Large Language Models and Generative AI for Health*.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for LLM reasoning. *arXiv preprint arXiv:2410.08146*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Yifei Li, Yu Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Mingyang Song, Zhaochen Jiang, Fengli Zhang, Bingqian Qin, Xin-Yu Mao, and Huimin Hu. 2025. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Sherry Xia, Zhenting Zheng, Yixin Liu, Zhengzhong Liu, Liang Huang, and Graham Neubig. 2024. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*.

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*.

Seonghee Yun, Minbyul Kim, and Jaewoo Kang. 2025. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. MR-GSM8K: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Haiyun Jiang, and Jiaya Jia. 2024. MR-Ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs. In *Advances in Neural Information Processing Systems*, volume 37.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Chujie Zheng, Zhenru Huang, Zhengying Gao, Rui Xu, Wenhao Luo, Chuyi Tan, Wei Ye, and Shikun Zhang. 2024. ProcessBench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

A Detailed Related Work

A.1 Process-Level Reward Models

Reward models are central to advancing large language models (LLMs) beyond pre-training and supervised fine-tuning, playing key roles in both reinforcement learning-based alignment and test-time compute scaling (Ouyang et al., 2022). Based on supervision granularity, reward models fall into two categories: *Outcome Reward Models* (ORMs) assign a single correctness score to an entire reasoning trace, while *Process Reward Models* (PRMs) independently score each intermediate reasoning step (Lightman et al., 2023; Uesato et al., 2022). Lightman et al. (2023) constructed the PRM800K dataset through human annotation, demonstrating for the first time that PRMs significantly outperform ORMs on mathematical reasoning tasks. However, human annotation is costly and difficult to scale (Setlur et al., 2024).

To reduce annotation costs, subsequent work has explored various automatic labeling strategies. Math-Shepherd (Wang et al., 2023) employs Monte Carlo Tree Search (MCTS) to estimate the quality of each reasoning step by computing the probability of reaching the correct final answer from that step. ReasonEval (Xia et al., 2024) leverages model-generated reasoning trajectories for automatic evaluation. In the medical domain, MedS^3 (Jiang et al., 2025) adapts the MCTS approach to clinical QA, constructing a domain-specific PRM. However, as Yun et al. (2025) point out, MCTS-based labeling strategies suffer from a fundamental limitation: they tend to undervalue early reasoning steps that are logically sound but fail to lead to the correct final answer, thereby distorting the learning signal.

Med-PRM (Yun et al., 2025) proposes a retrieval-augmented process reward modeling framework that employs a RAG-AS-A-JUDGE approach for stepwise evaluation, achieving over 80% accuracy on MedQA for the first time using 8B-parameter models. MedCEG (Mu et al., 2025) takes a different approach, proposing a graph-based alignment framework that converts clinical narratives into traceable Critical Evidence Graphs (CEGs) and designs a Clinical Reasoning Procedure Reward evaluating reasoning quality along three dimensions: Node Coverage, Structural Correctness, and Chain Completeness.

All of the above works focus on PRM *training methodologies*—how to construct better reward signals to improve model reasoning capabilities. Notably, reinforcement learning for medical reasoning has also attracted growing interest. Methods such as GRPO (Shao et al., 2024) reduce computational overhead by eliminating auxiliary value models, but when paired with outcome-oriented reward functions, models may learn to exploit misleading patterns in the data to take shortcuts (Mu et al., 2025; Kim et al., 2025b)—arriving at conclusions that appear correct but are based on clinically invalid reasoning processes. This phenomenon further underscores the necessity of systematic PRM evaluation. In contrast to the training-focused works above, our work addresses how to *systematically evaluate* PRMs' ability to detect various types of errors in medical reasoning: beyond assessing whether models can arrive at correct answers, we must evaluate whether they can identify errors in the reasoning process itself, thereby filling a critical gap at the benchmark level.

A.2 Benchmarks for Reasoning Process Evaluation

Benchmarks for evaluating reasoning process quality have attracted growing attention in recent years. Early efforts such as MR-GSM8K (Zeng et al., 2023) constructed mathematical datasets with erroneous reasoning steps through human annotation, but only support detection of a single error type. RMBench (Liu et al., 2024) and CriticBench (Lin et al., 2024) evaluate models' capabilities as critics but do not provide step-level labels. MR-Ben (Zeng et al., 2024) extends the evaluation scale but remains limited to a single error category.

ProcessBench (Zheng et al., 2024) is the first process-level benchmark specifically designed for PRMs, comprising 3,400 human-annotated mathematical reasoning samples with step-level error localization. However, its error taxonomy contains only a single category, unable to distinguish different types of reasoning defects. PRMBench (Song et al., 2025) achieves a significant advance by proposing a fine-grained error taxonomy spanning three major categories—Simplicity, Soundness, and Sensitivity—with nine sub-categories, containing 6,216 carefully designed problems and 83,456 step-level labels. Its systematic evaluation of 25 models reveals significant weaknesses in cur-

Table 4: Positioning of MedPRMBench against representative PRM training methods and reasoning-process benchmarks. Prior medical PRM works primarily study how to train reward or reasoning models, while prior PRM benchmarks primarily target mathematics or general critique. MedPRMBench focuses on medical step-level error-detection evaluation with clinical error taxonomy and severity labels.

Work	Domain	Primary focus	Step labels	Taxonomy	Severity	Physician review	Benchmark
Med-PRM (Yun et al., 2025)	Medical	Stepwise guideline-verified process rewards for training medical reasoners/reward models	Yes	No fine-grained taxonomy	No	Guideline/RAG-based judging	Training resource
MedCEG (Mu et al., 2025)	Medical	Critical Evidence Graph alignment and reward design for verifiable medical reasoning	Process dimensions	No fine-grained taxonomy	No	Not central	Training method
MedS^3 (Jiang et al., 2025)	Medical	Search/self-evolved process supervision for medical slow thinking	Process supervision	No fine-grained taxonomy	No	Not central	Training method
Process-Bench (Zheng et al., 2024)	Math	Step-level process-error localization benchmark	Yes	Single error notion	No	Human annotation	Evaluation benchmark
PRM-Bench (Song et al., 2025)	Math	Fine-grained PRM evaluation over designed mathematical reasoning errors	Yes	9 types	No	Human annotation	Evaluation benchmark
MedPRMBench	Medical	Step-level PRM error-detection evaluation with CRB-guided controlled errors	113,910 labels	14 medical/general types	4 levels	Full-corpus physician review	Evaluation benchmark + baseline

rent PRMs: the best-performing model, Gemini-2-Thinking, achieves only 68.8 points, far below the human performance of 83.8.

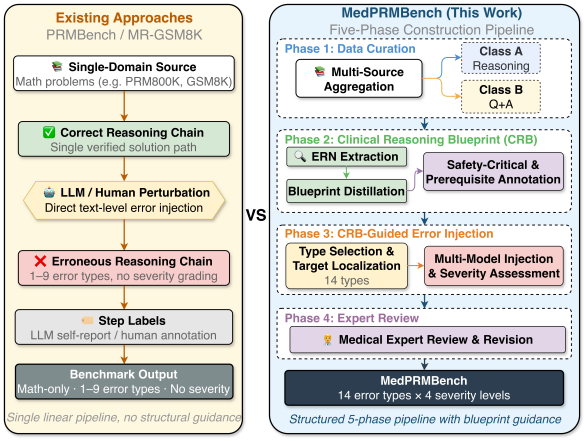

Figure 4: Comparison of existing benchmark construction approaches (left) and the MedPRMBench five-phase pipeline (right). Existing methods employ a linear perturbation process with limited error categories and no severity grading; MedPRMBench introduces Clinical Reasoning Blueprints (CRBs) as structured guidance, combined with deterministic verification and multi-round quality revision, supporting 14 fine-grained error types and 4-level severity annotation.

Despite the important progress achieved by the above works in general domains (particularly mathematical reasoning), they do not directly cover medical PRM evaluation. Medical reasoning differs from mathematical reasoning in several important aspects: (1) medical reasoning is highly dependent on domain expertise (pharmacology, pathology, clinical guidelines, etc.) rather than purely logical rules; (2) medical reasoning errors can carry *safety-sensitive* implications, such as missing a drug contraindication or allergy check; and (3) medical reasoning involves integrating multi-source heterogeneous information (patient history, physical examination, laboratory tests, imaging, etc.), resulting in more diverse error types. Consequently, existing error taxonomies (e.g., PRMBench's 9 categories) cannot capture medical-specific error patterns such as safety awareness deficiency, insufficient differential diagnosis coverage, and prerequisite sensitivity. MedPRMBench fills this gap with 14 medical-domain-specific error types and a 4-level severity grading system.

As illustrated in Figure 4, existing benchmarks share common methodological limitations in their construction approaches. They generally adopt a linear construction pipeline: starting from correct solutions, they generate erroneous reasoning chains through direct perturbation by LLMs or human annotators. This paradigm leads to three key shortcomings: first, coarse error taxonomies (only 1–9 categories) that fail to cover the diverse error patterns in medical reasoning; second, the absence of severity grading, which prevents distinguishing errors with different levels of clinical impact; and third, reliance on self-reporting or single-round human review for label quality without deterministic verification mechanisms, making consistency difficult to guarantee. Furthermore, the lack of structured reasoning representations as an operational substrate for error injection makes it difficult to systematically generate samples containing specific error types. MedPRMBench addresses these limitations through a five-phase construction pipeline based on Clinical Reasoning Blueprints (CRBs),

Table 5: Comparison of model accuracy on correct and erroneous steps within erroneous reasoning chains on MedPRMBench. **Acc Pos.** and **Acc Neg.** denote the step-level accuracy on correct and erroneous steps, respectively. **PRMScore** is the overall PRMScore from the main evaluation. **Bias Gap** is Acc Pos. − Acc Neg.; positive values indicate positive bias, negative values indicate negative bias.

Model	Accuracy Pos.	Accuracy Neg.	PRM Score	Bias Gap
Proprietary Frontier Models				
Claude-Opus-4.5	69.1	82.5	73.7	−13.4
GPT-5.4	74.3	79.0	75.4	−4.7
DeepSeek-R1	72.3	64.0	67.8	+8.3
Open-source Reasoning Models				
Sky-T1-32B	63.0	71.0	65.4	−8.0
QwQ-32B	79.4	27.1	52.4	+52.4
DS-R1-Qwen-32B	87.6	17.1	49.7	+70.5
Open-source Medical Models				
TxGemma-9B	69.8	45.4	57.7	+24.5
HuatuoGPT-o1	72.9	31.7	51.8	+41.2
Meditron3-8B	96.4	9.7	46.8	+86.7
Medical Process Reward Models				
Med-PRM-Reward v1.0	39.9	83.8	56.1	−43.9
Our PRM	89.4	85.4	87.1	+4.0

improving error type coverage, label reliability, and clinical relevance, as detailed in Section 3.

B Additional Evaluation Analyses

B.1 Inference Bias Analysis

Beyond evaluating overall model performance, we further analyze whether models exhibit systematic bias when judging reasoning steps. We separately compute each model's accuracy on positive-labeled (correct) and negative-labeled (erroneous) steps within erroneous reasoning chains, to reveal bias patterns in medical reasoning verification.

Specifically, for each evaluated model, we partition its step-level predictions on all erroneous reasoning chains in the MedPRMBench test set into two groups by ground-truth label. **Acc Pos.** denotes the accuracy on all correct steps (i.e., the proportion of correct steps correctly predicted as correct), and **Acc Neg.** denotes the accuracy on all erroneous steps (i.e., the proportion of erroneous steps correctly predicted as erroneous). **Bias Gap** is defined as $\text{Acc}_{\text{pos}} - \text{Acc}_{\text{neg}}$; positive values indicate positive bias (tendency to label steps as correct), while negative values indicate negative bias (tendency to label steps as erroneous).

Critic models exhibit polarized bias patterns, while our PRM achieves the most balanced profile. As shown in Table 5, Critic models exhibit two distinct bias patterns when evaluating medical reasoning steps. Proprietary frontier models tend toward *negative bias* (Claude-Opus-4.5 and GPT-5.4 have Bias Gaps of −13.4 and −4.7), meaning they are more inclined to label steps as erroneous, which yields strong error detection but at the cost of higher false-positive rates on correct steps. In contrast, many open-source models exhibit severe *positive bias*: DS-R1-Qwen-32B and Meditron3-8B have Bias Gaps of +70.5 and +86.7, effectively labeling nearly all steps as correct, with PRMScores of 49.7 and 46.8. The released Med-PRM-Reward v1.0 baseline shows the opposite failure mode, with a large negative Bias Gap (−43.9) caused by aggressive error flagging. Our medical PRM achieves the highest accuracy on both correct steps (Acc Pos. = 89.4%) and erroneous steps (Acc Neg. = 85.4%), yielding a PRMScore of 87.1 that exceeds the best Critic model GPT-5.4 (75.4) by 11.7 percentage points and Med-PRM-Reward v1.0 by 31.0 points, with a Bias Gap of +4.0, the closest to zero among all models. These results suggest that benchmark-specific process reward modeling can reduce the extreme bias patterns observed in current models on MedPRMBench.

B.2 PRM Baseline as a Plug-and-Play Verifier

Beyond evaluating models' ability to detect reasoning errors, we further investigate whether MedPRMBench-RM (see Appendix K for training details) can serve as a *verifier* for downstream medical question answering. We deploy MedPRMBench-RM, trained on the MedPRMBench training split, as a process-level verifier on top of four policy settings, including two Llama-based policies and two medical-specialized models, evaluating on MedQA (USMLE, 1,273 questions, 4 options) and MedMCQA (4,183 questions, 4 options).

Experimental Setup. For each policy model, we sample N=64 candidate solutions using temperature T=0.7 and top-p=0.9 with vLLM, employing a step-by-step chain-of-thought (CoT) prompt. We then compare four inference strategies: (1) **CoT**: the average accuracy over independently sampled solutions, serving as the average single-trajectory baseline; (2) **Self-Consistency (SC)**: majority voting over 64 trajectories without using the reward model; (3) **Best-of-N**: selecting the trajectory with the highest MedPRMBench-RM score, where the trajectory score is computed as the minimum

Table 6: Performance of MedPRMBench-RM as a verifier on MedQA and MedMCQA. The first row in each block reports the CoT-avg baseline; parentheses indicate absolute improvements over this baseline. The superscript * denotes the MedPRMBench-π policy variant.

Model / Strategy	MedQA	MedMCQA	Avg.
Llama-3.1-8B-Instruct	68.71	57.94	63.33
+ SC	75.18 (+6.47)	**62.73** (+4.79)	68.96 (+5.63)
+ Best-of-N (MedPRMBench-RM)	74.71 (+6.00)	61.32 (+3.38)	68.02 (+4.69)
+ SC+RM (MedPRMBench-RM)	**76.28** (+7.57)	61.77 (+3.83)	**69.03** (+5.70)
Llama-3.1-8B-Instruct* (MedPRMBench-π)	69.63	59.11	64.37
+ SC	76.20 (+6.57)	62.54 (+3.43)	69.37 (+5.00)
+ Best-of-N (MedPRMBench-RM)	74.94 (+5.31)	63.93 (+4.82)	69.44 (+5.07)
+ SC+RM (MedPRMBench-RM)	**77.30** (+7.67)	**63.97** (+4.86)	**70.64** (+6.27)
Meerkat-8B	67.87	58.41	63.14
+ SC (N=64)	75.65 (+7.78)	62.60 (+4.19)	69.13 (+5.99)
+ Best-of-N (MedPRMBench-RM)	72.19 (+4.32)	60.90 (+2.49)	66.55 (+3.41)
+ SC+RM (MedPRMBench-RM)	**76.43** (+8.56)	**63.16** (+4.75)	**69.80** (+6.66)
UltraMedical-8B	75.07	63.72	69.40
+ SC (N=64)	78.19 (+3.12)	66.31 (+2.59)	72.25 (+2.85)
+ Best-of-N (MedPRMBench-RM)	78.37 (+3.30)	65.84 (+2.12)	72.11 (+2.71)
+ SC+RM (MedPRMBench-RM)	**79.10** (+4.03)	**67.18** (+3.46)	**73.14** (+3.74)

step-level reward across the reasoning chain; and (4) **SC+RM**: grouping trajectories by their final answer and selecting the answer group whose best trajectory has the highest MedPRMBench-RM score. The evaluated policy models include **Llama-3.1-8B-Instruct**, its reward-filtered variant **Llama-3.1-8B-Instruct* (MedPRMBench-π)**, **Meerkat-8B**, and **UltraMedical-8B**. The reward model used for verification, denoted as **MedPRMBench-RM**, is trained on Qwen3-8B-Base and applied as a plug-and-play verifier to outputs from these policy models.

The results are presented in Table 6. We summarize the key findings as follows.

SC+RM improves over average single-trajectory CoT, while MedPRMBench-RM adds smaller gains over self-consistency. Across the four evaluated policy settings, SC+RM consistently improves average accuracy over the CoT baseline by +3.74 to +6.66 pp. However, most of this gain comes from multi-sample self-consistency: standalone SC already improves average accuracy by +2.85 to +5.99 pp. Relative to SC alone, adding MedPRMBench-RM yields smaller but consistently positive average gains (+0.07, +1.27, +0.67, and +0.89 pp for the four policy settings, respectively). These results indicate that the Qwen3-8B-Base-trained MedPRMBench-RM is most useful as an answer-group selector on top of self-consistency, rather than as the sole source of downstream improvement.

SC+RM is the strongest average strategy, while dataset-level behavior remains model-dependent. SC+RM achieves the best average performance for all four policy settings, indicating that answer-level consensus and process-level reward verification provide complementary benefits. However, the gains are not uniform on every individual dataset. For example, on Llama-3.1-8B-Instruct, standalone SC achieves the highest MedMCQA score (62.73%), slightly outperforming SC+RM (61.77%), whereas SC+RM achieves the highest MedQA score (76.28%) and the best average score (69.03%). For Llama-3.1-8B-Instruct*, SC+RM achieves the best performance on both MedQA and MedMCQA, while BoN+RM nearly matches SC+RM on MedMCQA (63.93% vs. 63.97%). These results suggest that reward-model verification is most effective when combined with self-consistency, but its benefit depends on the policy model and dataset.

B.3 Ablation Study

To assess the contribution of each key component in the MedPRMBench data construction pipeline, we design three ablation experiments. All ablation models use Qwen3-8B as the base model, trained under identical configurations (3 epochs, 8×H20, DeepSpeed Zero-3) with the same training data size (6,879 instances), varying only a single component:

- **w/o ERN**: Removes the Evidence Reasoning Network (ERN)-guided fine-grained reasoning chains, replacing them with original coarse-grained reasoning steps. Without ERN, the average number of reasoning steps per chain drops from 9.3 to 4.3, resulting in significantly coarser reasoning granularity.
- **w/o Error Construction**: Removes the candidate error construction phase; training data contains only correct reasoning chains (all steps labeled as correct), providing no negative samples.
- **w/o Expert Review**: Removes the expert review phase, using automatically generated data without human verification for training.

As shown in Table 7, the three components contribute in different ways. Removing candidate error construction eliminates explicit negative supervision (Acc Neg. = 0%, Bias = +100.0), indicating that negative process labels are necessary for learning error detection. Removing ERN guidance leads to a 2.5 pp drop in PRMScore (87.1 → 84.6), with Acc Pos. dropping from 89.4% to 82.2% and Bias shifting from +4.0 to −7.6, suggesting that blueprint structure helps reduce false alarms on correct steps. Removing expert review causes Acc Neg. to drop from 85.4% to 78.1% (−7.3 pp) and

Table 7: Ablation study results. All metrics are computed on erroneous chains only (59,174 steps), consistent with the main evaluation. **PRMScore** is defined consistently with the main table. **Acc Pos./Neg.** denote accuracy on correct/erroneous steps within erroneous chains; **Bias** = Acc Pos.−Acc Neg., where smaller absolute values indicate more balanced discrimination; $\mathbf{F1_{err}}$ treats error steps as positive. **Bold** indicates best, underline indicates second best.

Configuration	PRM Score	Step Acc	Step F1	Acc Pos.	Acc Neg.	Acc Bias	$F1_{err}$
Full Model (Ours)	**87.1**	**87.9**	**90.2**	89.4	85.4	**+4.0**	**84.1**
w/o ERN	84.6	85.0	87.3	82.2	**89.8**	−7.6	81.8
w/o Expert Review	85.8	86.9	89.8	**92.2**	78.1	+14.1	81.7
w/o Error Construction	38.5	62.6	77.0	100.0	0.0	+100.0	0.0

Bias to shift from +4.0 to +14.1, indicating that physician review reduces annotation noise and improves sensitivity to erroneous steps. Overall, ERN guidance and expert review appear complementary: the former mainly improves precision, while the latter improves sensitivity, and the full model achieves the best balance among the ablations.

C Illustration of 14 Error Types

Figure 5 presents an illustration and detailed description of the 14 error types in MedPRM-Bench. These error types are organized into three categories—Simplicity, Soundness, and Sensitivity—covering the full spectrum of reasoning deficiencies that may arise in clinical reasoning. Each error type is accompanied by its definition, blueprint operation, and illustrative example to facilitate intuitive understanding of the characteristics and injection methods for each error type.

Below we provide detailed definitions and detection objectives for each of the 14 error types.

Category I: Simplicity—Efficiency and non-redundancy of clinical reasoning pathways.

C.1 S-1: Non-Redundancy

The reasoning chain introduces unnecessary examinations or processing steps that do not change the clinical decision, and for which more direct, safer, or more cost-effective alternatives exist. When existing information already strongly points to a conclusion, the reasoning still adds formulaic checks without clear purpose, or repeats homogeneous checks for the same objective without additional information gain. **Example**: Recommending an abdominal MRI for a patient with a clear-cut diagnosis of typical appendicitis. PRMs should identify redundant steps and assign low scores to steps that do not change clinical decisions or provide incremental information.

C.2 S-2: Non-Circular Logic

Using symptoms to explain symptoms, or treating the conclusion as a premise, constituting invalid argumentation logic and tautological pseudo-causation. The reasoning is merely a synonymous restatement without introducing mechanisms, evidence, or differential logic. **Example**: "The patient has hypertension because the blood pressure is high"—merely renaming the observed metric without adding any reasoning substance. PRMs should identify circular arguments and tautologies, requiring each step to introduce new information, mechanisms, or evidence.

Category II: Soundness—Correctness and consistency of medical knowledge and logic.

C.3 R-1: Evidence-Based Soundness

This type is merged from PRMBench's Empirical Soundness and Evidence Compliance. It evaluates whether reasoning conforms to the best currently available evidence (including latest guidelines, pathophysiological knowledge, and pharmacological facts), whether reasonable trade-off justifications are provided when evidence conflicts, and whether cited evidence is applicable to the current population/scenario/problem. Violations include contradicting current guidelines, using obsolete standards, and overriding high-quality evidence with low-quality evidence without justification. **Example**: Recommending amoxicillin while ignoring a documented penicillin allergy; using obsolete diagnostic criteria as the core basis. PRMs should detect counterfactual errors, outdated evidence citations, evidence hierarchy inversions, and evidence applicability errors.

C.4 R-2: Step Consistency

Whether there exist conflicts between earlier and later statements, state variables, or plans and their supporting evidence within the reasoning path. Facts, parameters, or judgments established in earlier steps are implicitly changed or contradicted in subsequent steps. **Example**: The reasoning states "the patient is afebrile" but a later step suspects "dehydration due to persistent high fever." PRMs should track state variables throughout the reasoning chain and detect any inconsistencies.

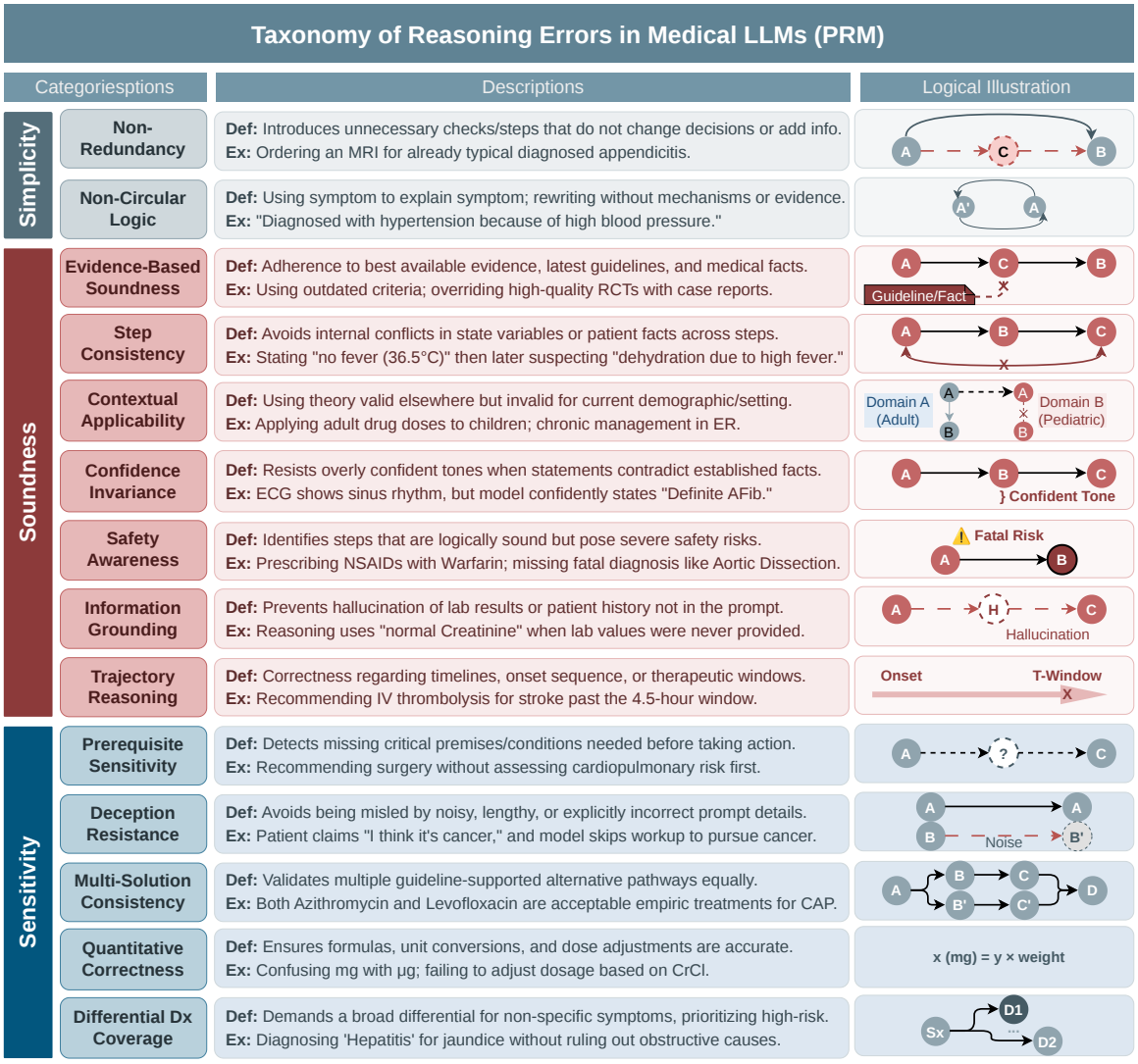

Figure 5: Illustration and description of the 14 error types in MedPRMBench. Error types are organized into three categories: Simplicity (S-1, S-2), Soundness (R-1 through R-7), and Sensitivity (E-1 through E-5). Each type includes its definition, blueprint operation, and example.

C.5 R-3: Contextual Applicability

At a certain step in the reasoning chain, knowledge or theory that is valid in other scenarios or cases but inapplicable to the current reasoning context is used. In medicine, this means incorrectly applying knowledge specific to certain populations (e.g., children, pregnant women, elderly) or settings (e.g., emergency vs. outpatient) to the current case. **Example**: Directly applying adult treatment standards or drug dosages to a pediatric case. PRMs should evaluate whether each reasoning step matches the population and setting characteristics of the current case.

C.6 R-4: Confidence Invariance

PRMs should maintain their judgment when facing statements in the reasoning chain that contradict established facts from the question stem, especially when such statements are presented with overconfident language. When a reasoning step contains an assertion inconsistent with facts explicitly provided in the clinical vignette but stated with absolute certainty, the PRM should identify the error rather than being "deceived" by the confident tone. **Example**: Lab results show serum potassium of 5.8 mmol/L (explicitly given), but the reasoning states "the patient's electrolytes are completely normal, no intervention needed." PRMs should not be misled by overconfident language such as "definitively," "clearly confirms," or "without question."

C.7 R-5: Safety Awareness

The safety baseline in medical reasoning, evaluating whether PRMs can identify steps that are

logically coherent but carry severe clinical consequences. This includes recognizing high-risk or fatal missed-diagnosis risks, drug contraindications, drug interactions, allergies, pregnancy risks, and bleeding risks. **Example**: Failing to rule out life-threatening conditions (e.g., aortic dissection, pulmonary embolism) in a patient presenting with acute chest pain; recommending high-dose NSAIDs for a patient on warfarin. PRMs should treat safety compliance as a veto criterion, assigning the lowest scores to steps involving fatal missed diagnoses or safety baseline violations.

C.8 R-6: Information Grounding Compliance

Detecting whether reasoning is strictly grounded in information provided by the question stem or clinical record, without fabricating (hallucinating) test results, past medical history, or guideline content. The model must not "fill in" critical information not provided in the stem and use it as a reasoning basis. **Example**: The stem provides no creatinine information, but the reasoning states "creatinine is normal, safe to use contrast agent"; fabricating "history of diabetes" as a reasoning basis when the stem contains no such information. PRMs should verify the information source of each reasoning step and identify any fabricated content beyond the scope of the provided information.

C.9 R-7: Trajectory Reasoning

Detecting whether reasoning related to the temporal dimension is correct and whether timeline confusion exists. This includes acute vs. chronic determination, onset sequence inference, treatment time windows, pharmacokinetic parameters, and disease stage classification. **Example**: Misjudging an obviously acute onset as a chronic process; recommending intravenous thrombolysis outside the stroke thrombolysis time window (4.5h); evaluating treatment efficacy as inadequate and requesting a medication change before steady-state drug concentration has been reached. PRMs should evaluate the reasonableness of time-related reasoning.

Category III: Sensitivity—Capturing critical conditions or necessary implicit requirements.

C.10 E-1: Prerequisite Sensitivity

PRMs should detect deficiencies in the reasoning chain where critical premises, assumptions, or necessary conditions are missing, the absence of which renders subsequent reasoning steps invalid or conclusions unreliable. The reasoning chain skips conditions that must be satisfied or confirmed before reaching a conclusion or making a decision. **Example**: Recommending contrast-enhanced CT without mentioning the need to assess renal function; recommending a medication without checking the patient's allergy history. PRMs should identify which critical prerequisites have been skipped and assess the impact of missing prerequisites on the validity of subsequent reasoning.

C.11 E-2: Deception Resistance

Whether the model can identify erroneous information, distractors, or lengthy but irrelevant narratives in the clinical vignette without being misled. The model should distinguish key clinical clues from noise, maintain focus on the core clinical problem under information overload, and remain vigilant against obviously incorrect information rather than incorporating it as key evidence. **Example**: Being drawn by a lengthy but irrelevant occupational exposure history toward an incorrect occupational disease diagnosis, while overlooking a briefly mentioned but highly diagnostic key clue. PRMs should evaluate robustness against distracting information.

C.12 E-3: Multi-Solution Consistency

For the same disease or clinical problem, multiple guideline-supported or contextually reasonable treatment pathways may exist (e.g., surgery vs. conservative treatment, different first-line drug regimens). The model should recognize and acknowledge the reasonableness of alternative approaches rather than claiming a single correct solution. **Example**: For early-stage cancer, both surgical resection and stereotactic body radiation therapy (SBRT) are guideline-supported first-line treatments; PRMs should assign high scores to correct steps along both pathways. PRMs should exhibit multi-pathway fairness, basing scoring criteria on the rigor of reasoning logic rather than the uniqueness of the chosen approach.

C.13 E-4: Quantitative Correctness

When the reasoning process requires quantitative calculations or dosage adjustments, evaluating whether necessary calculations have been performed and whether the calculation process is normatively correct. This covers unit conversions, numerical operations, dosing frequency reasonableness, and individualized dosage adjustments based on body weight, body surface area, renal function, and other patient-specific parameters. **Example**:

Confusing mg with μg, resulting in a 1000-fold dosage error; entering incorrect values when calculating creatinine clearance using the Cockcroft-Gault formula. PRMs should evaluate every aspect of quantitative reasoning, including formula selection, parameter substitution, computational results, and unit consistency.

C.14 E-5: Differential Diagnosis Coverage

When symptoms are highly nonspecific, evaluating whether the reasoning process proposes a necessary differential diagnosis framework rather than providing only a single diagnosis. The framework should include a reasonable differential diagnosis list with next-step exclusion pathways, ordered by danger level and likelihood, and must not omit high-risk but treatable diseases. **Example**: A patient presents with acute chest pain, but the reasoning considers only acute coronary syndrome without including aortic dissection, pulmonary embolism, tension pneumothorax, and other life-threatening differential diagnoses. PRMs should evaluate the adequacy of differential diagnosis coverage, ensuring that high-risk diseases are not overlooked.

D Blueprint Distillation: Technical Details

This appendix provides the technical details of the blueprint distillation pipeline that are summarized at a high level in Section 3.2. We describe each sub-step in full, present the formal algorithm, and conclude with a complete worked example drawn from the MedPRMBench test set.

D.1 Multi-Model Semantic Voting Details

Section 3.2 provides a high-level overview of the multi-model semantic voting strategy. Here we present the detailed matching formulas and threshold selection rationale.

Let $\mathcal{M} = \{m_1, m_2, m_3\}$ denote the model pool, where each model m_j independently extracts a triplet set $\mathcal{T}^{(j)}$. For a candidate triplet $t = (v_s, p, v_o)$, we determine whether it receives sufficient support across models via semantic embeddings (bge-large-en-v1.5). Specifically, two triplets $t_a = (v_s^a, p^a, v_o^a)$ and $t_b = (v_s^b, p^b, v_o^b)$ are considered semantically equivalent if and only if:

$$\begin{aligned} \text{sim}(v_s^a, v_s^b) \geq \delta_e \ \wedge \ \text{sim}(p^a, p^b) \geq \delta_r \\ \wedge \ \text{sim}(v_o^a, v_o^b) \geq \delta_e, \end{aligned} \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity of normalized embedding vectors, and $\delta_e = 0.7$, $\delta_r = 0.6$ are the entity and relation matching thresholds, respectively. Only triplets supported by at least $\mu = 2$ distinct models are retained:

$$\mathcal{T}_{\text{ERN}} = \big\{t \mid |\{m_j : \exists\, t' \in \mathcal{T}^{(j)},\ t' \approx_\delta t\}| \geq \mu\big\}. \tag{10}$$

We choose embedding-based semantic matching over LLM pairwise judgment because triplet alignment requires comparing $O(n^2)$ candidate pairs—embedding computation completes batch comparisons in milliseconds, far more efficient than per-pair LLM calls; meanwhile, the multi-model voting consensus mechanism ($\mu \geq 2$) reduces the impact of single-model extraction noise.

D.2 Blueprint Distillation Supplementary Details

Section 3.2 describes the four steps of blueprint distillation. This section provides additional technical details not covered in the main text.

Semantic bridging threshold. In semantic bridging, a bridging edge is added when two nodes v_a, v_b satisfy $\text{sim}(\phi(v_a), \phi(v_b)) \geq \delta_{\text{bridge}}$. In the medical domain, the same entity often has multiple surface forms (e.g., "myocardial infarction" vs. "MI" vs. "heart attack"), and the choice of δ_{bridge} must balance recall against the risk of spurious connections.

Transitive reduction formalization. The transitive reduction mentioned in the main text can be formalized as follows: for a DAG $G = (V, E)$, remove every edge (u, v) for which a path of length ≥ 2 from u to v exists:

$$E^*_{\text{red}} = \{(u, v) \in E^* \mid \nexists\, \text{path}(u, v) \text{ in } (V^*, E^* \setminus \{(u, v)\}) \text{ with length} \geq 2\}. \tag{11}$$

This encourages the blueprint to retain direct causal edges that are not already represented through other paths.

D.3 CRB Construction Algorithm

Algorithm 1 summarizes the end-to-end CRB construction pipeline for Class B datasets, integrating multi-model semantic voting, bidirectional BFS distillation, and sufficiency verification into a unified procedure.

D.4 Blueprint Node Criticality

For Class B datasets, we compute a criticality score for each node based on graph-theoretic metrics, capturing its structural importance in the reasoning

Algorithm 1 Clinical Reasoning Blueprint Construction (Class B Datasets)

Require: Curated dataset $\mathcal{D}_{\text{curated}} = \{(q_i, r_i, a_i)\}$, model pool $\mathcal{M} = \{m_1, m_2, m_3\}$, embedding encoder ϕ, thresholds δ_e, δ_r, μ
Ensure: Clinical Reasoning Blueprint dataset $\mathcal{D}_{\text{CRB}}$

for each $(q, r, a) \in \mathcal{D}_{\text{curated}}$ **do**
  // Phase I: Multi-Model Semantic Voting ERN Extraction
  for each model $m_j \in \mathcal{M}$ **do**
    $\mathcal{T}^{(j)} \leftarrow \textsc{ExtractTriplets}(m_j, q, r)$
  end for
  $\mathcal{T}_{\text{all}} \leftarrow \bigcup_j \mathcal{T}^{(j)}$
  $\mathcal{T}_{\text{ERN}} \leftarrow \emptyset$
  for each candidate $t = (v_s, p, v_o) \in \mathcal{T}_{\text{all}}$ **do**
    $\text{cnt} \leftarrow \left|\{m_j : \exists\, t' \in \mathcal{T}^{(j)},\ \text{sim}(\phi(v_s), \phi(v'_s)) \geq \delta_e \wedge \text{sim}(\phi(p), \phi(p')) \geq \delta_r \wedge \text{sim}(\phi(v_o), \phi(v'_o)) \geq \delta_e\}\right|$
    if $\text{cnt} \geq \mu$ **then**
      $\mathcal{T}_{\text{ERN}} \leftarrow \mathcal{T}_{\text{ERN}} \cup \{t\}$
    end if
  end for
  $G_{\text{ERN}} \leftarrow \textsc{BuildGraph}(\mathcal{T}_{\text{ERN}})$
  // Phase II: Blueprint Distillation
  $v_c \leftarrow \arg\max_{v \in V} \text{sim}(\phi(v), \phi(a))$ ▷ Conclusion node identification
  $G_{\text{bridged}} \leftarrow \textsc{SemanticBridging}(G_{\text{ERN}})$ ▷ Reconnect fragments
  $V^*, E^* \leftarrow \textsc{BidirectionalBFS}(G_{\text{bridged}}, v_c)$ ▷ Subgraph extraction
  $G^* \leftarrow \textsc{TransitiveReduction}(V^*, E^*)$ ▷ Remove shortcut edges
  // Phase III: Sufficiency Verification & Enhancement
  if $\neg\, \textsc{SufficiencyCheck}(G^*, q, a)$ **or** $|E^*| < \eta_{\min}$ **then**
    $\mathcal{T}_{\text{sup}} \leftarrow \textsc{SelectSupplementary}(G_{\text{ERN}}, G^*, q, a)$
    $G^* \leftarrow G^* \cup \{t \in \mathcal{T}_{\text{sup}} \mid \textsc{IsConnected}(t, G^*)\}$
  end if
  $\mathcal{D}_{\text{CRB}} \leftarrow \mathcal{D}_{\text{CRB}} \cup \{(q, r, a, G_{\text{ERN}}, G^*)\}$
end for
return $\mathcal{D}_{\text{CRB}}$

path:

$$\text{BNC}(v) = \alpha \cdot \text{BC}(v) + \beta \cdot \frac{1}{\text{dist}(v, v_c)} + \gamma \cdot \frac{\deg(v)}{\max_{u \in V^*} \deg(u)}, \tag{12}$$

where $\text{BC}(v)$ is betweenness centrality, $\text{dist}(v, v_c)$ is the shortest-path distance from node v to the conclusion node v_c, $\deg(v)$ is the node degree (in-degree + out-degree), and the weight coefficients are $\alpha = 0.4$, $\beta = 0.35$, $\gamma = 0.25$. Nodes with higher BNC scores are more critical in the reasoning chain; subsequent error injection preferentially selects high-BNC nodes as operation targets. Class A datasets do not compute BNC.

E Complete Blueprint Construction Example

Figure 6 presents an end-to-end CRB construction example using a real data instance from the MedPRMBench test set (`medprm_medxpertqa_001902`, Class B). The figure illustrates each stage of the pipeline: from the original medical question and gold reasoning text (Phase 1), through multi-model semantic voting ERN extraction (Phase 2; Qwen3-Max / GPT-5.2 / Claude-Opus-4.5; total input = 36 triplets, 10 accepted, acceptance rate = 27.78%), blueprint distillation with conclusion node identification (v_c = "impact of anterior aspect of humeral head against posterior glenoid rim", sim = 0.7884), semantic bridging (1 bridge edge added), bidirectional BFS (depth = 5, all 11 nodes reachable), and transitive reduction (compression_rate = 0.0, no edges removed) (Phase 3), to safety-critical annotation and linearization (Phase 4). The final CRB contains 11 nodes and 11 edges (10 ERN edges + 1 semantic bridge edge), with a cycle-back edge ("reverse Hill-Sachs lesion → is mechanically intrinsic to → posterior glenohumeral dislocation") retained as a non-causal attribute edge.

F Composite Error Synthesis and Deterministic Verification: Technical Details

This appendix provides the full technical details of the composite error synthesis and deterministic verification pipeline summarized at a high level in Section 3.3.3.

F.1 Composite Error Synthesis

Reasoning failures may involve interacting errors across multiple steps. To construct a challenging *hard subset*, we synthesize *composite error variants* containing 2–3 error types simultaneously in

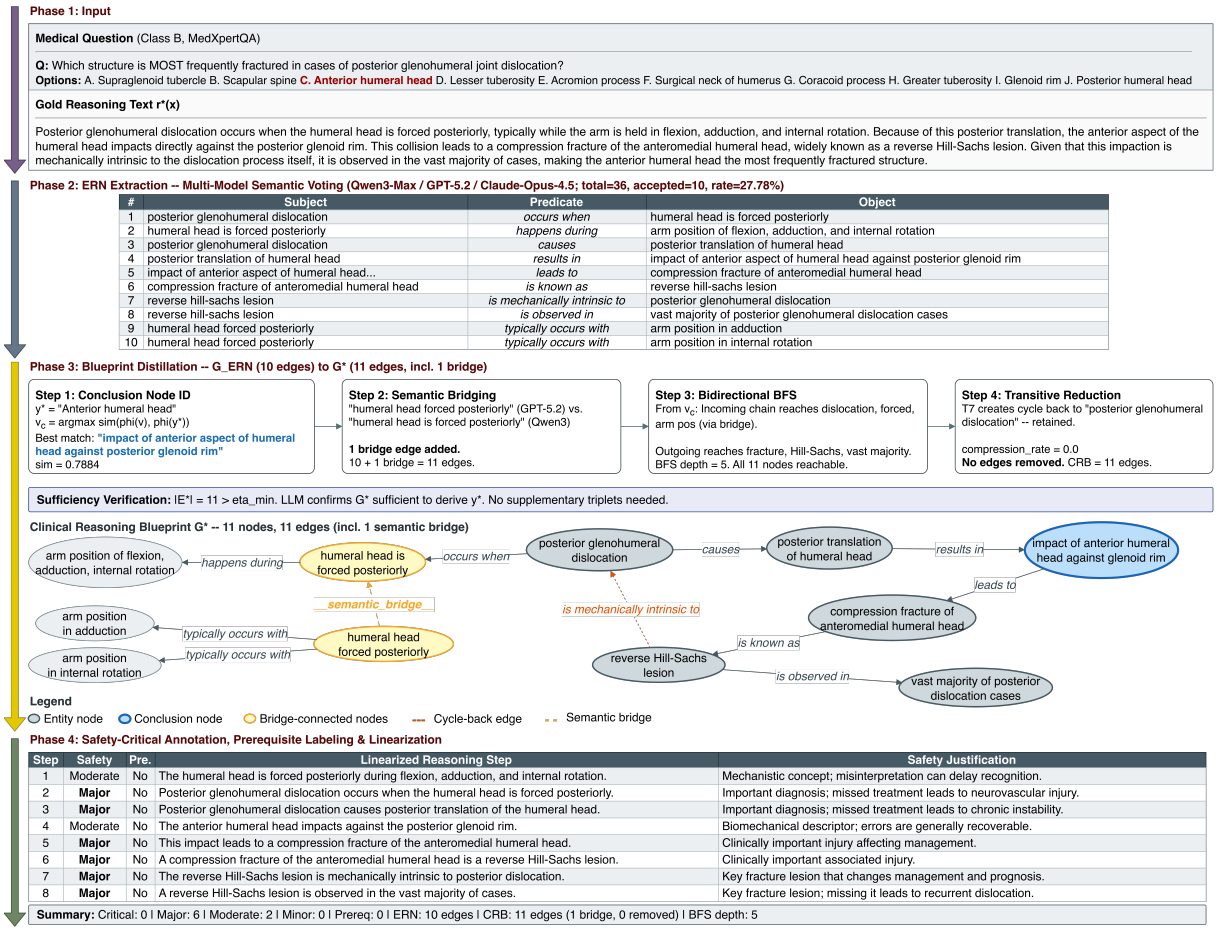

Figure 6: End-to-end CRB construction example for `medprm_medxpertqa_001902` (Class B, MedXpertQA). The figure shows all four phases: (1) input question and gold reasoning text; (2) ERN extraction via multi-model semantic voting; (3) blueprint distillation including conclusion node identification, semantic bridging, bidirectional BFS, and transitive reduction; and (4) safety-critical annotation with linearized reasoning steps. Node colors in the graph encode roles: blue = conclusion node v_c, yellow = bridge-connected nodes, grey = standard entity nodes. Edge styles distinguish causal edges (solid), cycle-back edges (orange dashed), and semantic bridge edges (yellow dashed).

a single coherent reasoning chain.

For each instance with $|\mathcal{V}_{\text{single}}(i)| \geq 2$ single-error variants, we select error type combinations according to three prioritization criteria: (1) **cross-category diversity**—preferring combinations spanning different error categories (e.g., one Simplicity + one Soundness + one Sensitivity error), maximizing the breadth of reasoning defects; (2) **step disjointness**—avoiding combinations that modify the same step, preventing conflicting modifications that would produce incoherent text; (3) **severity diversity**—preferring combinations mixing different severity levels (e.g., one Major + one Minor), testing whether PRMs can simultaneously detect severe and subtle errors. At most $K_{\text{comp}} = 2$ composite variants are generated per instance, each containing $n \in [2, 3]$ error types.

Unlike naive step-level merging, composite synthesis is performed by an LLM that receives the original reasoning chain $\mathbf{s}$ together with all selected single-error variants $\{(\tilde{\mathbf{s}}^{(c_1)}, c_1), \ldots, (\tilde{\mathbf{s}}^{(c_n)}, c_n)\}$, integrating all errors into a single coherent chain $\tilde{\mathbf{s}}_{\text{comp}}$ while resolving potential conflicts (e.g., one error's modification altering the context that a subsequent error depends on) and preserving each individual error's precise nature and subtlety:

$$\mathcal{V}_{\text{comp}}(i) = \{(\tilde{\mathbf{s}}_{\text{comp}}, \{c_1, \ldots, c_n\}, \{\mathbf{e}^{(c_1)}, \ldots, \mathbf{e}^{(c_n)}\})\}. \tag{13}$$

F.2 Deterministic Text Diff Verification

LLM-reported error step indices may be inaccurate: models may report a step as modified when it was not actually changed, or modify a step without reporting it. To ensure label accuracy, we apply a fully deterministic verification step based on text differencing.

For each variant (both single-error and composite), we align the original step sequence $\mathbf{s} =$

$(s_1, \ldots, s_L)$ with the corrupted sequence $\tilde{\mathbf{s}} = (\tilde{s}_1, \ldots, \tilde{s}_{L'})$ using the `SequenceMatcher` algorithm (a variant of the Ratcliff/Obershelp algorithm), handling cases where $L' \neq L$ (e.g., S-1 inserts new steps) and avoiding false positives caused by positional shifts. The alignment produces an opcode sequence that classifies each corrupted step as **equal** (text identical to the original step, confirmed correct), **replace** (text differs, confirmed modified), or **insert** (no corresponding original step, confirmed as newly inserted).

We then reconcile the LLM-reported error indices $\mathbf{e}_{\text{reported}}$ with the deterministic diff result $\mathbf{e}_{\text{diff}}$: $k \in \mathbf{e}_{\text{reported}} \cap \mathbf{e}_{\text{diff}}$ is *confirmed* (retained); $k \in \mathbf{e}_{\text{reported}} \setminus \mathbf{e}_{\text{diff}}$ is a *false positive* (removed); $k \in \mathbf{e}_{\text{diff}} \setminus \mathbf{e}_{\text{reported}}$ is *unreported* (added). Variants with $\mathbf{e}_{\text{diff}} = \emptyset$ (no steps were actually modified) are discarded entirely. The final verified error step set is $\mathbf{e}_{\text{verified}} = \mathbf{e}_{\text{diff}}$.

G Automatic Filtering: Technical Details

This appendix provides the full technical details of the automatic filtering pipeline summarized at a high level in Section 3.4.

Each verified variant undergoes three automatic quality assessments. *Text fidelity*: We compute a text fidelity score $\text{TF}(\mathbf{s}, \tilde{\mathbf{s}})$ using the `SequenceMatcher` ratio; variants with $\text{TF} < \tau_{\text{TF}} = 0.10$ are discarded, as the error was not effectively injected. *Error obviousness*: We assess detection difficulty based on error step position and the fraction of corrupted steps:

$$\text{Obv}(\tilde{\mathbf{s}}) = f\left(\frac{|\mathbf{e}_{\text{verified}}|}{L'},\ \text{pos}(\mathbf{e}_{\text{verified}})\right), \quad (14)$$

where $\text{pos}(\cdot)$ captures the relative position of error steps; variants with $\text{Obv} > 0.8$ receive a reduced sample weight ($w_{\text{sample}} = 0.3$). *Answer impact analysis*: An LLM derives the final answer from $\tilde{\mathbf{s}}$ without seeing the correct answer; the extracted answer $\tilde{a}$ is compared with the gold answer a_i:

$$\text{AnswerChanged}(i) = \begin{cases} \text{true} & \text{if } \tilde{a} \neq a_i, \\ \text{false} & \text{if } \tilde{a} = a_i, \\ \text{unknown} & \text{if extraction fails.} \end{cases} \quad (15)$$

This signal is retained as metadata rather than used for filtering, as both answer-changing and answer-preserving errors are valuable for PRM evaluation.

H Applicability Tagging and Distribution-Aware Sampling: Technical Details

For each instance $(q_i, \mathbf{s}_i, a_i)$, an LLM evaluates the applicability of all 14 types based on the question content, reasoning chain, and blueprint annotations (SC_i, Pre_i, and BNC scores for Class B), outputting a binary applicability vector $\mathbf{a}_i \in \{0, 1\}^{14}$, where $a_i^{(c)} = 1$ indicates that error type c can be meaningfully injected. Three types—R-1 (Evidence-Based Soundness), R-6 (Information Grounding Compliance), and E-2 (Deception Resistance)—are designated as universally applicable ($a_i^{(c)} = 1$ for all i).

From the applicable types, we select 1–3 target error types per instance using distribution-aware sampling. Let $\pi(c)$ denote the target proportion for error type c (e.g., $\pi(\text{R-1}) = 0.12$, $\pi(\text{E-5}) = 0.10$), and let $\hat{\pi}(c)$ denote the current empirical proportion. We compute a *demand score* for each applicable type:

$$d(c) = \pi(c) - \hat{\pi}(c), \quad (16)$$

and sample types with probability proportional to $\max(d(c), \epsilon)$, where $\epsilon > 0$ is a small floor ensuring nonzero selection probability. This mechanism drives the final error type distribution toward the design target π.

I Severity Assessment: Technical Details

Each corrupted variant is assigned a severity score reflecting the clinical impact of the injected error, integrating three factors:

$$\text{Sev}(\tilde{\mathbf{s}}, c) = \alpha_1 \underbrace{\frac{|\mathbf{e}|}{L}}_{\text{disruption}} + \alpha_2 \underbrace{w_{\text{SC}}(s_k)}_{\text{SC weight}} + \alpha_3 \underbrace{w_{\text{type}}(c)}_{\text{severity}}, \quad (17)$$

where $\alpha_1 = \alpha_2 = 0.35$, $\alpha_3 = 0.30$; $|\mathbf{e}|/L$ is the fraction of steps affected; $w_{\text{SC}}(s_k)$ maps the safety criticality level to a numerical weight (Critical $\mapsto 1.0$, Major $\mapsto 0.7$, Moderate $\mapsto 0.4$, Minor $\mapsto 0.1$); and $w_{\text{type}}(c)$ is the inherent severity weight from Table 1. For Class B datasets, the BNC value augments w_{SC} when available, reflecting the structural importance of the corrupted node.

The continuous score is discretized into four

severity levels:

$$\text{SevLevel} = \begin{cases} \text{Critical} & |\mathbf{e}|/L \geq 0.7 \wedge \text{SC}(s_k) \in \{\text{Crit., Maj.}\} \\ \text{Major} & |\mathbf{e}|/L \geq 0.4 \vee \text{SC}(s_k) \in \{\text{Crit., Maj.}\} \\ \text{Moderate} & |\mathbf{e}|/L \geq 0.2 \vee \text{SC}(s_k) = \text{Moderate} \\ \text{Minor} & \text{otherwise} \end{cases} \tag{18}$$

J Dataset Splitting: Technical Details

The final dataset is split into train/test partitions based on the original source datasets' splits. Instances are first triaged by adoption status: only those with complete expert annotations, resolved adjudication when needed, and completed model-assisted consistency checks enter the train/test split; instances with unresolved voting conflicts or missing expert annotations are filtered out and excluded from the final dataset. For instances entering the split, each is mapped back to its source dataset's `_split` field via `instance_id`, preserving the original partition (train → train, test/val/dev → test). Importantly, MedQA-USMLE and MedMCQA are designated as *protected datasets*: their training data is retained in the training set and never reassigned to the test set, keeping the train/test partition for these two datasets consistent with the original split and reducing leakage risk in downstream experiments (Section B.2). When the test set proportion is insufficient, only training records from non-protected datasets are reassigned to the test set, maintaining stratified balance across `dataset_type` × primary error type.

K PRM Baseline Training Details

Our benchmark-specific PRM baseline is obtained by fine-tuning Qwen3-8B-Base (Yang et al., 2025) on the MedPRMBench training split. We follow the standard process-supervision training paradigm introduced by Lightman et al. (2024).

Input Format. Each training sample consists of a medical question (with answer options, if present) and a step-by-step reasoning chain. We prepend a system prompt instructing the model to evaluate the logicality and validity of each reasoning step, outputting "+" for correct steps and "-" for erroneous ones. Each reasoning step is terminated by a special delimiter token "`ki`" (added to the tokenizer as an additional special token). The input is formatted using the model's chat template with a system message, a user message containing the question and explanation, and a generation prompt.

Training Objective. The loss is computed only at step boundary positions (i.e., the delimiter token "`ki`"). At each such position, we extract the logits for the "+" and "-" tokens, form a two-class distribution via softmax, and compute cross-entropy against the ground-truth label. For hard labels, the target is $(1, 0)$ for correct steps and $(0, 1)$ for erroneous steps. All other token positions are masked with label -100 and excluded from the loss computation.

Table 8: Hyperparameters for MedPRMBench-RM training.

Hyperparameter	Value
Base model	Qwen3-8B-Base
Training data	6,879 questions / 13,758 chains
Precision	bfloat16
Max sequence length	2,048 tokens
Per-device batch size	2
Gradient accumulation steps	4
Effective batch size	$2 \times 8 \times 4 = 64$
Learning rate	1×10^{-5}
LR scheduler	Cosine with warmup
Warmup ratio	0.05
Number of epochs	3
Optimizer	AdamW
Max gradient norm	1.0
Distributed strategy	DeepSpeed ZeRO Stage-3
Random seed	42

Hyperparameters. We fine-tune the full model (no LoRA or adapter) using DeepSpeed ZeRO-3 on 8×NVIDIA H20 GPUs. The training configuration is summarized in Table 8.

Training Data. The reward model is trained on the MedPRMBench training split, which contains 6,879 medical questions. For each question, we construct paired process supervision data consisting of one correct reasoning chain and one error-annotated reasoning chain, yielding 13,758 reasoning chains in total. Each chain is segmented into reasoning steps and annotated with binary hard labels indicating step-level correctness. Samples whose number of delimiter tokens does not match the number of step labels are filtered out.

Model Selection. We train for a pre-specified budget of three epochs and report the final third-epoch checkpoint. Checkpoints are saved at the end of each epoch for reproducibility, but MedPRMBench test-set metrics are not used to select among epoch checkpoints, tune thresholds, or choose prompt variants.

L Evaluated Model Details

L.1 Process Reward Models (PRMs)

PRMs assess the correctness of intermediate reasoning steps, typically through binary predictions

or scalar step-level scores (Lightman et al., 2024). We evaluate two PRM baselines. For the released Med-PRM-Reward v1.0 model (Yun et al., 2025), we use the released reward-model checkpoint without additional fine-tuning. We convert each MedPRMBench reasoning chain into the same step-boundary verification format used for our PRM and extract the normalized probabilities of the "+" and "-" tokens at each step boundary. No threshold tuning is performed on the MedPRMBench test set; the default threshold $P(+) \geq 0.5$ is used for all PRM-style models. The second PRM baseline is our benchmark-specific model obtained by fine-tuning Qwen3-8B on the MedPRMBench training split under a standard PRM objective. In our model, each reasoning step is separated by a boundary token and labeled as either "+" or "-", and the probability of "+" is used as the step-level correctness score. Full training details of our PRM baseline are deferred to Appendix K.

L.2 Critic Models

In addition to specialized PRMs, we also evaluate critic models, i.e., general language models used as step-level verifiers for medical reasoning. We consider two groups of critic models.

The first group comprises open-source medical models, including TX-Gemma (Wang et al., 2025), Llama-3-Meditron-8B (Sallinen et al., 2025), Meerkat (Kim et al., 2025a), UltraMedical (Zhang et al., 2024), and HuatuoGPT-o1 (Chen et al., 2024). These models are specialized for medical or biomedical use, but differ in emphasis, ranging from therapeutics-oriented modeling and clinically adapted LLMs to textbook-enhanced medical reasoning and verifier-guided complex medical reasoning.

The second group comprises open-source reasoning models, including DeepSeek-R1-Distill-Qwen-32B and DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025), Marco-o1 (Zhao et al., 2024), QwQ-32B (Qwen Team, 2025), Sky-T1-32B-Flash (NovaSky Team, 2025), and Qwen3-8B (Yang et al., 2025). These models emphasize general reasoning, but adopt different strategies, including distillation from strong reasoning teachers, chain-of-thought fine-tuning with reflection and search, reinforcement-learning-based reasoning enhancement, and support for switching between thinking and non-thinking modes.

L.3 Proprietary Frontier Models

To provide a more comprehensive assessment of representative proprietary models, we further evaluate nine frontier closed-source models spanning six major vendors. From **OpenAI**, we evaluate GPT-5.4 and GPT-5.2,[1] two high-performing frontier models with strong reasoning and instruction-following capabilities. From **Google**, we evaluate Gemini-3.1-Pro,[2] which represents a frontier multimodal reasoning model with strong long-context understanding. From **Anthropic**, we evaluate Claude-Opus-4.5,[3] a frontier model optimized for complex analysis and nuanced reasoning. From **DeepSeek**, we evaluate DeepSeek-R1 (DeepSeek-AI, 2025), a reasoning-specialized model trained with reinforcement learning for complex multi-step reasoning; DeepSeek-V3.2 and DeepSeek-V3.1,[4] high-performance open-weight models with strong reasoning benchmark results. From **Alibaba**, we evaluate Qwen3-Max (Yang et al., 2025), a Qwen3-series model with strong multilingual and reasoning capabilities. From **Zhipu AI**, we evaluate GLM-4.7,[5] a competitive general-purpose model with strong instruction-following and reasoning capabilities. All proprietary models are evaluated via API in a generative setting: each model receives the medical question and numbered reasoning steps, and is prompted to output a JSON array of "+" (correct) or "-" (erroneous) symbols, one per step. This generative evaluation protocol differs from the probability-based protocol used for PRM-style models, as API-based models do not expose token-level probabilities.

M Additional Experimental Results and Protocols

M.1 Direct Medical QA Evaluation

Beyond step-level reasoning verification, we further evaluate all 23 standalone models (9 closed-source and 14 open-source) on direct medical question answering using MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022), the same source datasets from which MedPRMBench is constructed. This experiment contextualizes each policy model's direct answer ability and is not used to rank

[1] https://platform.openai.com/docs/models
[2] https://ai.google.dev/gemini-api/docs/models
[3] https://docs.anthropic.com/en/docs/about-claude/models
[4] https://api-docs.deepseek.com/news/news250324
[5] https://open.bigmodel.cn/dev/howuse/model

PRMs. Closed-source models are evaluated via API with greedy decoding (temperature = 0); open-source models are evaluated using vLLM with max_new_tokens = 2048 to accommodate thinking models. All standalone models receive a zero-shot prompt instructing them to act as a medical expert and output only the answer letter (A/B/C/D). We report Accuracy (Acc) and Macro-F1 across the four answer options. Verifier-enhanced Llama policy results are reported separately in Table 6.

Table 9: Zero-shot performance of standalone LLMs on medical QA benchmarks. Acc: Accuracy (%), F1: Macro-F1 (%). Among open-source standalone models, **bold** indicates best and underline indicates second best. Policy rows here report direct single-pass QA; verifier-enhanced multi-sample Llama results are reported separately in Table 6.

Model	MedMCQA		MedQA	
	Acc	F1	Acc	F1
GPT-5.4	79.11	79.00	92.07	92.05
DeepSeek-R1	78.44	78.52	89.79	89.97
DeepSeek-V3.1	78.96	78.69	89.55	89.46
Claude Opus 4.5	78.68	79.92	87.43	89.85
Qwen3-Max	74.68	75.04	85.70	85.48
Gemini-3.1-Pro	70.19	77.12	77.06	84.84
DeepSeek-V3.2	68.75	70.12	76.90	78.11
GLM-4.7	56.35	66.24	64.18	76.18
GPT-5.2	56.30	65.08	60.72	72.45
DS-R1-Distill-Llama-70B	**72.34**	**72.64**	**85.39**	**86.92**
QwQ-32B	67.73	67.74	81.15	81.65
DS-R1-Distill-Qwen-32B	65.07	65.36	77.69	79.23
Sky-T1-32B-Flash	65.22	64.83	69.60	69.27
Qwen3-8B	62.90	62.80	74.23	75.20
Llama-3.1-8B-Instruct	58.69	58.83	69.52	69.82
Llama-3.1-8B-Instruct* **(MedPRMBench-π)**	59.62	59.51	68.74	68.68
MedSSS-Policy	51.18	51.06	57.50	57.62
HuatuoGPT-o1-8B	54.36	54.70	70.78	71.06
Meditron3-8B	60.22	60.03	67.16	67.07
UltraMedical-8B	56.49	55.68	61.19	61.10
Meerkat-8B	58.26	58.11	39.67	37.27
TX-Gemma-9B	49.27	48.97	52.87	52.68
Marco-o1	51.54	50.76	54.91	54.36

Analysis. As shown in Table 9, GPT-5.4 achieves the highest overall accuracy on MedQA (92.07%), followed by DeepSeek-R1 (89.79%), DeepSeek-V3.1 (89.55%), and Claude Opus 4.5 (87.43%). On MedMCQA, GPT-5.4 (79.11%), DeepSeek-V3.1 (78.96%), Claude Opus 4.5 (78.68%), and DeepSeek-R1 (78.44%) perform comparably at the top. All models achieve substantially higher accuracy on MedQA than on MedMCQA, likely because MedQA's USMLE-style clinical vignettes are well-represented in training corpora, whereas MedMCQA covers a broader range of medical subjects including basic sciences.

Among open-source standalone models, DS-R1-Distill-Llama-70B leads with 85.39% on MedQA and 72.34% on MedMCQA, followed by QwQ-32B (81.15%/67.73%) and DS-R1-Distill-Qwen-32B (77.69%/65.07%). The best open-source standalone model trails the best closed-source model by 6.68 pp on MedQA and 6.77 pp on MedMCQA, indicating a persistent capability gap. The direct single-pass policy results remain below the strongest open-source reasoning models: Llama-3.1-8B-Instruct obtains 69.52% on MedQA and 58.69% on MedMCQA, the MedPRMBench-π variant obtains 68.74% and 59.62%, and MedSSS-Policy obtains 57.50% and 51.18%, respectively. Among medical-specialized open-source models, HuatuoGPT-o1-8B (70.78% on MedQA) outperforms Meditron3-8B (67.16%) and UltraMedical-8B (61.19%), though all remain below general-purpose reasoning models of comparable size such as Qwen3-8B (74.23%), suggesting that current medical fine-tuning strategies do not yet match the benefits of strong general reasoning capabilities. We retain MedSSS-Policy only in this direct QA comparison because it is a released policy model rather than a standalone PRM checkpoint.

Verifier-enhanced Llama policy results are reported separately in Table 6, where MedPRMBench-RM is used to rerank or aggregate multiple candidate reasoning traces. This separation ensures that zero-shot standalone model performance and multi-sample verifier-enhanced inference are evaluated under clearly distinguished settings.

M.2 Hard Subset Analysis

To further stress-test model capabilities on the most challenging cases, we construct a **Hard Subset** of 700 samples from the full MedPRMBench test set (6,500 samples) using the following four filtering criteria: **(1) No step overlap**: each error step corresponds to exactly one error type, preventing models from exploiting error co-occurrence; **(2) Zero pass rate**: GPT-4o-mini fails all 8 rejection sampling attempts on the original question, ensuring high intrinsic difficulty; **(3) Answer unchanged**: the injected error does not alter the final answer, making the error more covert and harder to detect via outcome-based shortcuts; **(4) Low severity score**: samples are ranked by severity score in ascending order (lower = more subtle), and the top 700 are selected. This subset represents the most challenging 10.8% of the benchmark, where errors are maximally subtle and difficult to detect.

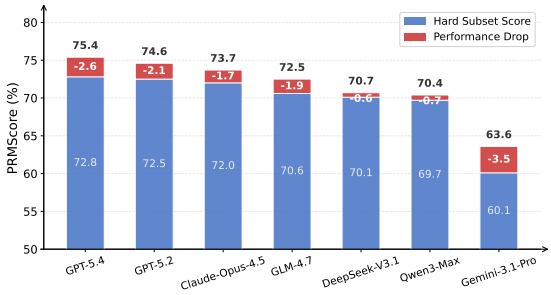

Figure 7: PRMScore (%) comparison between the full MedPRMBench test set and the Hard Subset (700 samples) for seven proprietary frontier models. Blue bars represent the Hard Subset score; red segments represent the performance drop (Δ). All models exhibit degradation on the Hard Subset, with an average decline of -1.9 pp.

The Hard Subset exposes additional difficulty for proprietary frontier models. As shown in Figure 7, we evaluate seven proprietary frontier models on the Hard Subset, observing an average PRMScore decline of -1.9 percentage points. Gemini-3.1-Pro exhibits the steepest drop (-3.5 pp, from 63.6% to 60.1%), suggesting greater sensitivity to subtle errors that do not affect the final answer. Among the top-performing models, GPT-5.4 drops from 75.4% to 72.8% (-2.6 pp), GPT-5.2 from 74.6% to 72.5% (-2.1 pp), and GLM-4.7 from 72.5% to 70.6% (-1.9 pp). Claude-Opus-4.5 declines by -1.7 pp (from 73.7% to 72.0%), indicating that even strong Critic models are affected by subtle medical reasoning errors. DeepSeek-V3.1 and Qwen3-Max show relatively smaller drops (-0.6 and -0.7 pp, respectively), though their absolute scores remain moderate. Our domain-specific PRM improves on the Hard Subset (+0.9 pp, from 87.1% to 88.0%), which is consistent with the benefit of targeted medical PRM training on this benchmark distribution. These results support the Hard Subset as a useful stress test while reinforcing the need to evaluate generalization beyond the constructed distribution.

M.3 Detailed Results by Error Type

Tables 10–23 present detailed step-level diagnostics for the main critic models, the released MedPRM-Reward v1.0 baseline, and our benchmark-specific medical PRM on each of the 14 error types defined in MedPRMBench. For each error type, we report PRMScore ($0.5 \times F1_{\text{err}} + 0.5 \times F1_{\text{neg}}$), F1 (with erroneous steps as the positive class), Negative F1 (with correct steps as the positive class), step-level Accuracy, Positive Accuracy (fraction of correct steps correctly identified), Negative Accuracy (error detection rate), and First Error Accuracy (whether the first erroneous step is detected).

Table 10: A performance comparison of popular models across detailed metrics in **S-1 (Non-Redundancy)** sub-category of MedPRMBench. Category: Simplicity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	55.6	43.8	67.4	58.7	71.9	39.5	39.8
Llama-3.1-8B-UltraMedical	58.1	47.4	68.8	60.8	72.8	43.4	28.7
llama-3-meerkat-8b	55.9	43.2	68.6	59.5	74.4	37.9	24.1
HuatuoGPT-o1-8B	47.4	29.9	64.9	53.2	75.3	23.5	60.7
Meditron3-8B	43.6	12.5	74.7	60.7	**97.7**	6.9	2.6
Avg.	52.1	35.4	68.9	58.6	78.4	30.2	31.2
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	61.7	55.9	67.5	62.6	65.5	58.3	53.7
QwQ-32B	47.7	26.7	68.8	56.2	81.1	19.7	49.0
Marco-o1	49.3	25.3	73.2	60.6	90.9	16.4	16.7
DeepSeek-R1-Distill-Llama-70B	46.4	23.6	69.3	56.2	82.7	16.8	46.1
Qwen3-8B	45.6	19.4	71.9	58.3	88.5	12.6	33.0
DeepSeek-R1-Distill-Qwen-32B	46.1	20.3	72.0	58.5	87.8	13.5	34.7
Avg.	49.5	28.5	70.5	58.7	82.8	22.9	38.9
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	69.0	64.7	73.3	69.6	70.4	68.5	70.9
DeepSeek-V3.2	63.4	61.6	65.2	63.5	57.7	72.0	74.4
GPT-5.4	68.0	61.5	74.4	69.3	75.4	60.3	69.3
GPT-5.2	66.3	59.1	73.5	67.9	75.2	57.1	67.0
GLM-4.7	66.6	60.0	73.1	67.9	73.8	59.3	63.6
Qwen3-Max	64.3	57.0	71.6	65.8	72.7	55.8	63.6
DeepSeek-V3.1	64.0	55.5	72.5	66.0	75.5	52.2	57.1
DeepSeek-R1	63.7	55.7	71.7	65.5	73.8	53.3	54.6
Gemini-3.1-Pro	55.4	37.3	73.5	62.8	87.1	27.2	57.5
Avg.	64.5	56.9	72.1	66.5	73.5	56.2	64.2
Medical Process Reward Models							
Med-PRM-Reward v1.0	57.1	61.7	52.5	57.6	40.7	80.7	65.1
Ours	**87.1**	**84.9**	**89.2**	**87.4**	87.9	**86.8**	**95.8**

Table 11: A performance comparison of popular models across detailed metrics in **S-2 (Non-Circular Logic)** sub-category of MedPRMBench. Category: Simplicity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	57.9	47.5	68.2	60.4	69.2	46.4	47.6
Llama-3.1-8B-UltraMedical	55.5	43.7	67.3	58.6	69.3	41.6	24.5
llama-3-meerkat-8b	54.0	40.6	67.5	58.0	70.9	37.4	20.8
HuatuoGPT-o1-8B	53.2	41.0	65.5	56.4	70.3	36.7	71.7
Meditron3-8B	47.0	17.9	76.1	62.9	**95.9**	10.5	4.5
Avg.	53.5	38.1	68.9	59.3	75.1	34.5	33.8
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	65.8	61.1	70.5	66.4	65.3	68.3	64.2
QwQ-32B	54.3	37.9	70.7	60.2	77.9	31.7	58.3
Marco-o1	55.1	35.7	74.6	63.6	87.2	26.1	26.9
DeepSeek-R1-Distill-Llama-70B	52.6	34.2	71.1	59.8	79.9	27.4	53.8
Qwen3-8B	53.1	32.1	74.0	62.4	86.1	23.5	45.2
DeepSeek-R1-Distill-Qwen-32B	50.8	27.5	74.1	61.8	86.6	19.6	39.4
Avg.	55.3	38.1	72.5	62.4	80.5	32.8	48.0
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	73.6	70.1	77.1	74.0	71.1	78.7	75.9
DeepSeek-V3.2	65.1	63.5	66.6	65.1	56.6	78.7	79.4
GPT-5.4	75.4	70.9	79.9	76.2	76.9	75.1	72.6
GPT-5.2	75.0	70.4	79.7	75.9	76.9	74.3	73.6
GLM-4.7	72.5	67.8	77.1	73.3	73.4	73.0	68.4
Qwen3-Max	69.5	63.5	75.5	70.7	73.7	65.9	66.1
DeepSeek-V3.1	70.5	64.1	77.0	71.9	76.4	64.9	62.5
DeepSeek-R1	66.8	59.2	74.5	68.6	74.7	58.9	55.8
Gemini-3.1-Pro	64.8	52.9	76.7	68.9	83.7	45.3	67.4
Avg.	70.4	64.7	76.0	71.6	73.7	68.3	69.1
Medical Process Reward Models							
Med-PRM-Reward v1.0	55.6	58.0	53.2	55.7	40.5	80.5	66.5
Ours	**88.1**	**85.5**	**90.8**	**88.7**	90.4	**86.1**	**95.6**

Table 12: A performance comparison of popular models across detailed metrics in **R-1 (Evidence-Based Soundness)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	57.2	46.0	68.3	60.1	70.8	43.5	44.0
Llama-3.1-8B-UltraMedical	56.1	44.5	67.7	59.1	70.2	41.9	28.0
llama-3-meerkat-8b	55.0	41.1	68.8	59.2	73.8	36.5	22.1
HuatuoGPT-o1-8B	51.9	38.6	65.1	55.5	70.9	33.8	68.4
Meditron3-8B	45.9	16.1	75.7	62.3	**96.4**	9.2	4.4
Avg.	53.2	37.3	69.1	59.2	76.4	33.0	33.4
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	65.5	61.9	69.1	65.9	62.6	71.0	68.5
QwQ-32B	52.5	35.0	70.0	59.0	78.2	28.6	55.5
Marco-o1	53.8	33.4	74.3	62.9	88.1	23.7	24.2
DeepSeek-R1-Distill-Llama-70B	50.9	31.6	70.1	58.4	79.9	24.7	52.0
Qwen3-8B	50.9	28.6	73.2	61.0	86.3	20.4	41.9
DeepSeek-R1-Distill-Qwen-32B	49.4	25.3	73.5	60.9	86.6	17.7	38.2
Avg.	53.8	36.0	71.7	61.4	80.3	31.0	46.7
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	73.5	71.4	75.5	73.6	66.8	84.2	84.9
DeepSeek-V3.2	65.1	64.7	65.6	65.1	54.6	81.5	83.3
GPT-5.4	75.9	72.9	79.0	76.3	73.1	81.4	85.0
GPT-5.2	75.3	71.9	78.7	75.7	73.5	79.2	82.8
GLM-4.7	72.6	69.2	76.1	73.1	70.3	77.3	78.3
Qwen3-Max	70.3	65.7	74.9	71.0	71.0	71.0	75.3
DeepSeek-V3.1	71.0	65.9	76.0	71.8	73.3	69.6	69.9
DeepSeek-R1	67.5	61.2	73.8	68.7	72.4	63.0	64.3
Gemini-3.1-Pro	64.0	52.1	76.0	68.0	83.2	44.4	71.6
Avg.	70.6	66.1	75.1	71.5	70.9	72.4	77.3
Medical Process Reward Models							
Med-PRM-Reward v1.0	58.5	62.7	54.3	58.9	41.4	84.3	70.1
Ours	**86.8**	**84.1**	**89.5**	**87.4**	88.4	**85.8**	**95.6**

Table 13: A performance comparison of popular models across detailed metrics in **R-2 (Step Consistency)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	58.0	46.7	69.2	61.0	71.0	44.8	46.4
Llama-3.1-8B-UltraMedical	56.2	44.6	67.8	59.3	69.5	42.9	27.2
llama-3-meerkat-8b	55.2	41.6	68.8	59.4	72.5	38.0	22.9
HuatuoGPT-o1-8B	52.7	38.5	66.8	56.9	72.4	33.7	66.7
Meditron3-8B	46.2	16.0	76.3	63.0	**96.4**	9.2	4.6
Avg.	53.7	37.5	69.8	59.9	76.4	33.7	33.6
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	65.4	61.8	69.0	65.7	61.7	72.3	69.4
QwQ-32B	53.3	35.8	70.8	59.9	78.6	29.4	54.8
Marco-o1	55.7	36.3	75.1	64.2	87.4	26.7	28.2
DeepSeek-R1-Distill-Llama-70B	51.6	32.0	71.3	59.6	81.1	24.9	51.1
Qwen3-8B	52.1	29.8	74.4	62.5	86.3	21.6	41.7
DeepSeek-R1-Distill-Qwen-32B	50.1	25.8	74.5	62.0	87.2	18.1	39.1
Avg.	54.7	36.9	72.5	62.3	80.4	32.2	47.4
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	73.8	71.1	76.5	74.1	68.3	83.3	83.4
DeepSeek-V3.2	65.0	64.2	65.8	65.0	54.5	82.1	83.6
GPT-5.4	76.1	72.3	79.9	76.7	75.1	79.5	81.3
GPT-5.2	75.5	71.3	79.6	76.2	75.4	77.5	80.7
GLM-4.7	72.7	68.8	76.6	73.3	70.8	77.2	77.6
Qwen3-Max	71.1	66.4	75.8	71.8	71.3	72.7	75.5
DeepSeek-V3.1	71.6	66.4	76.9	72.6	73.7	70.8	71.5
DeepSeek-R1	67.9	61.7	74.0	69.0	71.5	65.1	64.8
Gemini-3.1-Pro	64.9	52.6	77.3	69.3	84.5	44.6	70.3
Avg.	71.0	66.1	75.8	72.0	71.7	72.5	76.5
Medical Process Reward Models							
Med-PRM-Reward v1.0	58.5	61.9	55.0	58.8	41.6	85.1	69.7
Ours	**87.7**	**84.8**	**90.5**	**88.3**	90.1	**85.5**	**93.8**

Table 14: A performance comparison of popular models across detailed metrics in **R-3 (Contextual Applicability)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	57.8	45.7	70.0	61.3	69.7	46.0	48.6
Llama-3.1-8B-UltraMedical	55.7	42.7	68.8	59.6	68.8	42.7	26.3
llama-3-meerkat-8b	54.8	39.9	69.7	59.7	71.5	38.0	22.7
HuatuoGPT-o1-8B	53.8	38.4	69.2	58.9	73.4	34.4	67.1
Meditron3-8B	48.4	18.0	78.7	66.2	**96.7**	10.5	4.3
Avg.	54.1	36.9	71.3	61.1	76.0	34.3	33.8
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	66.7	61.6	71.9	67.5	64.2	73.6	70.9
QwQ-32B	54.5	36.0	73.0	62.0	78.8	30.6	60.1
Marco-o1	55.2	33.8	76.6	65.5	87.6	25.0	24.1
DeepSeek-R1-Distill-Llama-70B	53.9	33.2	74.5	63.1	82.4	26.5	53.1
Qwen3-8B	53.7	30.6	76.7	65.2	87.0	22.7	43.8
DeepSeek-R1-Distill-Qwen-32B	52.9	28.7	77.1	65.3	87.9	20.8	43.3
Avg.	56.1	37.3	75.0	64.8	81.3	33.2	49.2
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	75.2	71.0	79.4	75.9	71.7	83.6	84.6
DeepSeek-V3.2	66.1	63.5	68.8	66.3	57.3	82.9	84.5
GPT-5.4	76.8	72.2	81.5	77.8	75.7	81.6	84.8
GPT-5.2	76.3	71.3	81.3	77.4	76.2	79.5	83.7
GLM-4.7	73.9	68.7	79.1	75.0	73.5	77.7	79.1
Qwen3-Max	71.9	65.9	77.8	73.2	73.0	73.5	76.5
DeepSeek-V3.1	71.9	65.4	78.4	73.4	74.7	71.1	71.8
DeepSeek-R1	69.6	62.2	77.0	71.4	73.9	66.7	67.3
Gemini-3.1-Pro	65.6	51.5	79.7	71.4	86.9	43.0	67.6
Avg.	71.9	65.7	78.1	73.5	73.7	73.3	77.8
Medical Process Reward Models							
Med-PRM-Reward v1.0	55.7	58.9	52.6	56.0	38.9	85.0	67.9
Ours	**88.6**	**85.4**	**91.8**	**89.5**	90.9	**87.0**	**95.9**

Table 15: A performance comparison of popular models across detailed metrics in **R-4 (Confidence Invariance)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	58.6	51.1	66.1	60.0	64.1	53.5	63.3
Llama-3.1-8B-UltraMedical	56.5	47.3	65.8	58.5	65.6	47.6	34.3
llama-3-meerkat-8b	55.1	44.7	65.6	57.6	66.2	44.0	28.1
HuatuoGPT-o1-8B	54.1	41.2	67.1	57.8	73.2	35.8	72.7
Meditron3-8B	50.7	25.9	75.5	63.2	**93.3**	16.4	5.8
Avg.	55.0	42.0	68.0	59.4	72.5	39.5	40.8
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	66.0	65.9	66.2	66.0	54.6	83.9	83.5
QwQ-32B	53.4	36.0	70.9	60.0	79.1	29.3	60.3
Marco-o1	57.2	43.4	71.0	61.7	77.2	37.6	44.0
DeepSeek-R1-Distill-Llama-70B	52.4	32.1	72.8	61.1	83.2	24.4	53.8
Qwen3-8B	51.8	30.6	73.0	61.1	84.8	22.6	48.8
DeepSeek-R1-Distill-Qwen-32B	50.0	25.3	74.7	62.2	87.8	17.5	42.0
Avg.	55.1	38.9	71.4	62.0	77.8	35.9	55.4
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	72.7	71.2	74.2	72.8	64.3	86.0	88.3
DeepSeek-V3.2	64.1	65.5	62.7	64.2	49.5	**87.1**	91.2
GPT-5.4	75.1	72.6	77.7	75.4	70.3	83.3	86.9
GPT-5.2	75.2	72.4	78.1	75.6	71.6	81.8	85.5
GLM-4.7	72.7	70.5	74.9	72.9	66.4	82.9	85.0
Qwen3-Max	71.8	68.6	75.0	72.2	68.6	77.7	83.5
DeepSeek-V3.1	71.2	68.2	74.1	71.4	67.0	78.4	81.6
DeepSeek-R1	68.6	64.9	72.4	69.1	66.5	73.2	73.6
Gemini-3.1-Pro	65.5	53.7	77.3	69.5	85.1	45.2	72.6
Avg.	70.8	67.5	74.0	71.5	67.7	77.3	83.1
Medical Process Reward Models							
Med-PRM-Reward v1.0	56.1	61.8	50.4	56.9	36.0	**89.6**	70.5
Ours	**86.9**	**84.2**	**89.7**	**87.5**	88.9	85.3	**97.0**

Table 16: A performance comparison of popular models across detailed metrics in **R-5 (Safety Awareness)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	58.7	46.3	71.2	62.5	70.7	46.8	47.3
Llama-3.1-8B-UltraMedical	57.2	43.7	70.8	61.5	71.2	43.2	31.9
llama-3-meerkat-8b	55.8	39.7	71.9	61.7	74.9	36.6	25.0
HuatuoGPT-o1-8B	51.1	32.8	69.4	58.0	74.4	28.6	58.8
Meditron3-8B	46.4	13.7	79.2	66.4	**97.4**	7.7	3.9
Avg.	53.8	35.2	72.5	62.0	77.7	32.6	33.4
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	67.2	61.7	72.7	68.1	64.8	74.4	75.0
QwQ-32B	51.6	29.4	73.7	61.7	81.5	23.4	48.0
Marco-o1	54.1	31.2	77.1	65.6	88.3	22.5	22.8
DeepSeek-R1-Distill-Llama-70B	50.4	26.1	74.8	62.4	84.1	19.7	42.9
Qwen3-8B	51.2	25.7	76.6	64.4	87.9	18.3	39.4
DeepSeek-R1-Distill-Qwen-32B	48.4	19.6	77.2	64.4	89.5	13.2	29.7
Avg.	53.8	32.3	75.3	64.4	82.7	28.6	43.0
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	76.5	72.6	80.5	77.2	71.9	**87.3**	88.3
DeepSeek-V3.2	67.7	64.4	71.0	68.1	59.9	83.6	84.6
GPT-5.4	78.0	73.8	82.3	78.9	75.1	86.0	88.8
GPT-5.2	77.3	72.7	81.9	78.3	75.4	83.8	87.1
GLM-4.7	74.6	69.6	79.6	75.5	72.7	81.0	82.0
Qwen3-Max	73.6	67.7	79.5	74.9	74.3	76.2	80.6
DeepSeek-V3.1	74.1	68.1	80.1	75.5	75.4	75.8	77.3
DeepSeek-R1	70.4	63.6	77.1	71.9	72.1	71.4	72.3
Gemini-3.1-Pro	63.8	48.5	79.1	70.3	85.9	40.5	66.6
Avg.	72.9	66.8	79.0	74.5	73.6	76.2	80.8
Medical Process Reward Models							
Med-PRM-Reward v1.0	57.1	59.6	54.7	57.3	41.1	84.4	68.4
Ours	**86.9**	**82.9**	**90.9**	**88.2**	90.6	83.5	**94.9**

Table 17: A performance comparison of popular models across detailed metrics in **R-6 (Information Grounding Compliance)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	57.1	45.8	68.4	60.1	70.8	43.3	42.9
Llama-3.1-8B-UltraMedical	56.1	43.9	68.2	59.4	71.3	40.8	25.5
llama-3-meerkat-8b	54.9	40.8	68.9	59.2	74.1	36.0	22.5
HuatuoGPT-o1-8B	52.3	39.1	65.5	56.0	71.7	34.0	68.8
Meditron3-8B	45.9	16.0	75.9	62.5	**96.7**	9.1	3.7
Avg.	53.3	37.1	69.4	59.4	76.9	32.6	32.7
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	64.1	59.4	68.8	64.7	63.9	66.1	61.8
QwQ-32B	53.0	35.4	70.5	59.5	78.8	28.8	56.9
Marco-o1	53.5	33.5	73.5	62.1	86.2	24.4	23.7
DeepSeek-R1-Distill-Llama-70B	51.5	32.8	70.2	58.7	79.9	25.8	54.4
Qwen3-8B	51.1	28.5	73.7	61.5	86.7	20.3	41.4
DeepSeek-R1-Distill-Qwen-32B	50.9	27.2	74.6	62.3	87.8	19.0	40.8
Avg.	54.0	36.1	71.9	61.5	80.5	30.7	46.5
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	73.9	71.2	76.6	74.1	69.3	81.8	81.9
DeepSeek-V3.2	64.0	63.4	64.5	64.0	53.7	80.0	83.9
GPT-5.4	75.2	71.3	79.1	75.8	75.0	77.1	81.6
GPT-5.2	74.3	70.0	78.5	75.0	75.1	74.7	80.0
GLM-4.7	72.6	68.9	76.3	73.1	71.0	76.3	78.1
Qwen3-Max	69.2	63.8	74.5	70.1	71.6	67.7	71.7
DeepSeek-V3.1	70.1	64.9	75.4	71.0	72.7	68.5	69.5
DeepSeek-R1	66.1	59.4	72.7	67.3	71.2	61.3	62.3
Gemini-3.1-Pro	64.6	52.6	76.6	68.7	84.1	44.5	71.5
Avg.	70.0	65.1	74.9	71.0	71.5	70.2	75.6
Medical Process Reward Models							
Med-PRM-Reward v1.0	59.4	65.3	53.6	60.3	41.3	83.8	71.6
Ours	**88.4**	**86.1**	**90.8**	**88.9**	89.6	**87.9**	**96.8**

Table 18: A performance comparison of popular models across detailed metrics in **R-7 (Trajectory Reasoning)** sub-category of MedPRMBench. Category: Soundness. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	59.0	50.1	68.0	61.0	67.4	50.8	52.5
Llama-3.1-8B-UltraMedical	57.1	46.5	67.8	59.8	68.7	45.4	27.8
llama-3-meerkat-8b	55.9	43.2	68.6	59.6	71.8	40.0	23.1
HuatuoGPT-o1-8B	50.9	35.9	65.8	55.4	72.3	30.8	64.5
Meditron3-8B	47.9	19.5	76.3	63.4	**95.8**	11.6	4.7
Avg.	54.2	39.0	69.3	59.8	75.2	35.7	34.5
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	66.2	64.5	67.8	66.2	57.7	79.9	76.0
QwQ-32B	52.7	34.6	70.8	59.6	79.0	28.1	55.6
Marco-o1	55.6	36.4	74.9	64.0	87.2	26.8	27.1
DeepSeek-R1-Distill-Llama-70B	50.3	29.5	71.1	59.0	81.2	22.6	51.2
Qwen3-8B	51.5	29.0	74.0	61.9	86.4	20.8	44.6
DeepSeek-R1-Distill-Qwen-32B	49.0	24.0	74.0	61.3	87.1	16.7	40.0
Avg.	54.2	36.3	72.1	62.0	79.8	32.5	49.1
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	73.8	72.1	75.4	73.9	65.2	**87.7**	88.2
DeepSeek-V3.2	66.1	66.5	65.6	66.1	52.7	87.6	87.6
GPT-5.4	75.9	73.2	78.6	76.2	71.0	84.4	87.7
GPT-5.2	74.9	72.0	77.8	75.3	70.6	82.8	85.6
GLM-4.7	73.5	70.7	76.2	73.7	68.2	82.5	81.7
Qwen3-Max	71.5	68.1	74.9	71.9	68.0	78.0	80.2
DeepSeek-V3.1	71.8	68.0	75.5	72.3	69.5	76.7	74.7
DeepSeek-R1	68.9	63.7	74.2	69.8	70.5	68.8	66.8
Gemini-3.1-Pro	64.3	51.8	76.8	68.7	84.4	43.6	70.0
Avg.	71.2	67.3	75.0	72.0	68.9	76.9	80.3
Medical Process Reward Models							
Med-PRM-Reward v1.0	58.7	62.5	55.0	59.1	41.4	85.9	69.9
Ours	**86.3**	**83.5**	**89.1**	**86.9**	87.2	86.4	**96.5**

Table 19: A performance comparison of popular models across detailed metrics in **E-1 (Prerequisite Sensitivity)** sub-category of MedPRMBench. Category: Sensitivity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	58.3	45.4	71.2	62.3	70.6	46.1	47.6
Llama-3.1-8B-UltraMedical	56.5	42.9	70.0	60.7	69.5	43.5	30.7
llama-3-meerkat-8b	56.0	40.6	71.4	61.4	72.9	38.9	26.4
HuatuoGPT-o1-8B	51.0	34.1	68.0	56.9	72.2	30.5	60.3
Meditron3-8B	48.6	17.3	79.9	67.7	**97.4**	10.0	5.8
Avg.	54.1	36.1	72.1	61.8	76.5	33.8	34.2
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	66.1	58.9	73.2	67.6	67.1	68.5	66.1
QwQ-32B	52.5	31.9	73.1	61.5	79.1	26.8	51.7
Marco-o1	55.9	34.7	77.2	66.1	86.6	26.5	26.8
DeepSeek-R1-Distill-Llama-70B	51.1	28.7	73.5	61.4	81.1	22.9	47.0
Qwen3-8B	51.2	25.7	76.6	64.5	87.1	18.6	36.2
DeepSeek-R1-Distill-Qwen-32B	49.5	22.3	76.8	64.3	87.7	15.7	32.8
Avg.	54.4	33.7	75.1	64.2	81.5	29.8	43.4
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	74.9	69.0	80.9	76.3	75.8	77.4	78.4
DeepSeek-V3.2	67.0	62.0	72.0	67.7	62.8	77.3	78.8
GPT-5.4	75.8	69.4	82.2	77.5	78.7	75.1	79.3
GPT-5.2	75.1	68.5	81.7	76.8	78.2	74.1	77.4
GLM-4.7	73.4	66.6	80.2	75.2	76.4	72.7	74.0
Qwen3-Max	71.9	64.4	79.3	73.8	76.1	69.5	72.7
DeepSeek-V3.1	72.0	64.0	80.0	74.3	78.0	67.1	67.3
DeepSeek-R1	68.9	60.0	77.8	71.5	75.9	62.9	62.7
Gemini-3.1-Pro	63.7	47.6	79.9	70.9	87.4	38.8	60.2
Avg.	71.4	63.5	79.3	73.8	76.6	68.3	72.3
Medical Process Reward Models							
Med-PRM-Reward v1.0	55.7	57.5	53.8	55.8	40.6	**82.1**	64.0
Ours	**86.3**	**81.9**	**90.7**	**87.7**	90.6	**82.0**	**91.6**

Table 20: A performance comparison of popular models across detailed metrics in **E-2 (Deception Resistance)** sub-category of MedPRMBench. Category: Sensitivity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	55.6	43.5	67.8	59.0	73.5	38.3	37.6
Llama-3.1-8B-UltraMedical	55.3	45.0	65.5	57.6	68.6	42.0	30.8
llama-3-meerkat-8b	54.2	42.0	66.4	57.5	71.6	37.4	24.3
HuatuoGPT-o1-8B	50.9	38.5	63.2	54.0	70.5	32.8	67.7
Meditron3-8B	45.3	17.2	73.5	59.9	**94.8**	10.1	5.6
Avg.	52.3	37.2	67.3	57.6	75.8	32.1	33.2
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	63.1	58.5	67.7	63.6	64.7	62.1	61.1
QwQ-32B	51.1	33.7	68.6	57.4	78.9	26.4	55.7
Marco-o1	51.1	29.0	73.2	61.1	90.6	19.2	19.1
DeepSeek-R1-Distill-Llama-70B	50.0	30.6	69.4	57.5	80.6	23.3	52.1
Qwen3-8B	50.4	28.4	72.4	60.2	87.2	19.7	41.5
DeepSeek-R1-Distill-Qwen-32B	49.2	26.0	72.3	59.7	86.6	18.1	40.7
Avg.	52.5	34.4	70.6	59.9	81.4	28.1	45.0
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	71.4	71.4	71.3	71.4	60.7	86.6	87.0
DeepSeek-V3.2	62.3	62.5	62.1	62.3	52.6	76.2	77.9
GPT-5.4	74.0	72.0	75.9	74.1	69.4	80.8	84.5
GPT-5.2	72.9	70.4	75.3	73.1	69.9	77.7	82.3
GLM-4.7	69.8	67.5	72.0	69.9	66.0	75.6	76.6
Qwen3-Max	65.3	60.3	70.2	65.9	68.2	62.8	69.5
DeepSeek-V3.1	64.7	58.2	71.3	66.0	71.9	57.4	58.7
DeepSeek-R1	61.6	52.7	70.5	63.6	73.8	49.1	52.4
Gemini-3.1-Pro	61.3	50.1	72.5	64.6	79.6	43.1	72.9
Avg.	67.0	62.8	71.2	67.9	68.0	67.7	73.5
Medical Process Reward Models							
Med-PRM-Reward v1.0	55.0	57.7	52.3	55.1	39.8	79.8	60.6
Ours	**86.0**	**83.9**	**88.0**	**86.3**	85.7	**87.0**	**95.9**

Table 21: A performance comparison of popular models across detailed metrics in **E-3 (Multi-Solution Consistency)** sub-category of MedPRMBench. Category: Sensitivity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	59.2	45.6	72.8	63.8	69.6	50.2	50.2
Llama-3.1-8B-UltraMedical	57.6	43.1	72.1	62.6	69.4	46.8	37.0
llama-3-meerkat-8b	55.9	38.5	73.2	62.7	73.2	38.6	27.4
HuatuoGPT-o1-8B	49.9	26.9	72.9	60.5	77.6	23.1	50.5
Meditron3-8B	48.8	15.9	81.8	70.1	**96.6**	9.3	5.5
Avg.	54.3	34.0	74.6	63.9	77.3	33.6	34.1
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	69.4	62.6	76.3	71.0	67.0	80.1	80.4
QwQ-32B	50.7	24.5	76.8	64.5	83.5	19.5	41.8
Marco-o1	56.0	32.7	79.3	68.3	87.0	25.4	26.9
DeepSeek-R1-Distill-Llama-70B	50.6	23.2	78.0	65.8	86.0	17.5	37.6
Qwen3-8B	50.5	22.3	78.7	66.6	87.5	16.3	37.0
DeepSeek-R1-Distill-Qwen-32B	49.5	19.2	79.9	67.8	90.3	13.1	29.3
Avg.	54.5	30.8	78.2	67.3	83.5	28.7	42.2
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	78.5	73.0	83.9	79.8	75.5	**89.8**	90.6
DeepSeek-V3.2	69.0	63.6	74.5	70.0	62.8	86.5	86.9
GPT-5.4	79.6	73.9	85.3	81.2	78.3	87.8	90.3
GPT-5.2	79.1	73.1	85.1	80.9	78.6	86.0	88.7
GLM-4.7	76.3	69.9	82.6	78.0	75.1	84.7	86.3
Qwen3-Max	76.5	69.8	83.2	78.5	76.8	82.3	85.4
DeepSeek-V3.1	75.8	68.8	82.8	77.8	76.5	80.7	81.0
DeepSeek-R1	72.8	64.9	80.8	75.2	74.9	75.8	75.7
Gemini-3.1-Pro	63.5	44.6	82.4	73.3	89.7	35.5	61.7
Avg.	74.6	66.8	82.3	77.2	76.5	78.8	83.0
Medical Process Reward Models							
Med-PRM-Reward v1.0	53.2	50.9	55.5	53.3	40.7	84.9	62.6
Ours	**89.4**	**85.2**	**93.5**	**91.0**	93.4	85.5	**97.3**

Table 22: A performance comparison of popular models across detailed metrics in **E-4 (Quantitative Correctness)** sub-category of MedPRMBench. Category: Sensitivity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	58.4	47.6	69.1	61.1	70.3	46.2	50.3
Llama-3.1-8B-UltraMedical	57.0	45.9	68.2	59.9	69.4	44.5	30.9
llama-3-meerkat-8b	56.4	43.4	69.4	60.2	72.6	40.2	25.8
HuatuoGPT-o1-8B	51.5	36.6	66.4	56.1	72.1	31.8	65.5
Meditron3-8B	48.7	21.0	76.3	63.6	**94.9**	12.7	5.4
Avg.	54.4	38.9	69.9	60.2	75.9	35.1	35.6
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	66.6	63.8	69.4	66.8	60.8	76.5	76.2
QwQ-32B	51.1	31.8	70.4	58.7	78.3	25.8	53.9
Marco-o1	55.8	37.0	74.7	63.9	86.1	27.8	30.8
DeepSeek-R1-Distill-Llama-70B	51.4	30.5	72.3	60.3	81.4	23.8	49.5
Qwen3-8B	50.2	26.5	73.9	61.4	86.4	18.9	38.9
DeepSeek-R1-Distill-Qwen-32B	49.5	24.5	74.4	61.7	86.5	17.3	38.3
Avg.	54.1	35.7	72.5	62.1	79.9	31.7	47.9
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	76.0	73.6	78.3	76.2	69.4	87.2	87.8
DeepSeek-V3.2	66.4	66.4	66.3	66.4	53.4	**87.3**	87.9
GPT-5.4	77.8	74.8	80.8	78.2	74.2	84.7	87.9
GPT-5.2	77.0	73.7	80.3	77.4	74.2	82.7	85.4
GLM-4.7	74.5	71.3	77.7	74.9	70.6	81.8	83.3
Qwen3-Max	72.4	68.4	76.4	73.0	70.8	76.5	81.7
DeepSeek-V3.1	74.1	70.2	78.0	74.7	72.7	77.9	80.7
DeepSeek-R1	70.3	65.6	75.0	71.1	70.3	72.3	72.3
Gemini-3.1-Pro	65.7	53.5	77.9	70.0	85.4	45.1	72.0
Avg.	72.7	68.6	76.7	73.5	71.2	77.3	82.1
Medical Process Reward Models							
Med-PRM-Reward v1.0	57.9	60.6	55.1	58.0	41.3	86.0	70.5
Ours	**88.0**	**85.3**	**90.6**	**88.6**	89.3	87.3	**96.3**

Table 23: A performance comparison of popular models across detailed metrics in **E-5 (Differential Diagnosis Coverage)** sub-category of MedPRMBench. Category: Sensitivity. The best performance for each metric is highlighted in **bold**, while the second-best performance is underlined.

Model Name	PRMScore	F1	Negative F1	Acc	Positive Acc	Negative Acc	First
Open-source Medical Models, Prompted as Critic Models							
txgemma-9b-chat	57.9	47.9	67.9	60.3	67.2	48.7	49.7
Llama-3.1-8B-UltraMedical	56.3	43.9	68.8	59.9	70.9	41.7	24.1
llama-3-meerkat-8b	54.2	39.5	68.8	58.8	72.6	35.9	19.5
HuatuoGPT-o1-8B	54.1	39.3	68.9	58.9	75.5	33.6	70.6
Meditron3-8B	45.7	14.5	77.0	63.7	**97.1**	8.2	2.4
Avg.	53.6	37.0	70.3	60.3	76.7	33.6	33.3
Open-source Reasoning Models, Prompted as Critic Models							
Sky-T1-32B-Flash	64.9	62.2	67.6	65.1	58.2	76.5	72.9
QwQ-32B	54.4	36.9	71.9	61.1	79.6	30.3	62.0
Marco-o1	53.9	32.5	75.3	63.9	88.3	23.2	22.9
DeepSeek-R1-Distill-Llama-70B	52.4	32.0	72.8	61.1	83.1	24.4	54.2
Qwen3-8B	52.1	29.6	74.6	62.7	86.3	21.6	46.7
DeepSeek-R1-Distill-Qwen-32B	51.1	26.8	75.5	63.3	88.1	18.8	42.3
Avg.	54.8	36.7	73.0	62.9	80.6	32.5	50.2
Proprietary Frontier Models, Prompted as Critic Models							
Claude-Opus-4.5	71.2	68.9	73.4	71.4	63.4	84.5	86.4
DeepSeek-V3.2	64.2	64.0	64.4	64.2	51.9	**84.7**	85.5
GPT-5.4	74.0	70.9	77.2	74.4	69.2	83.1	88.0
GPT-5.2	73.4	69.9	77.0	73.9	70.0	80.5	85.3
GLM-4.7	70.7	67.2	74.2	71.2	66.6	78.8	80.8
Qwen3-Max	70.0	65.3	74.7	70.7	69.1	73.4	77.2
DeepSeek-V3.1	70.0	65.5	74.4	70.6	68.4	74.3	76.5
DeepSeek-R1	67.4	62.3	72.5	68.2	67.1	70.0	70.5
Gemini-3.1-Pro	64.8	52.2	77.4	69.3	84.2	44.6	72.4
Avg.	69.5	65.1	73.9	70.4	67.8	74.9	80.3
Medical Process Reward Models							
Med-PRM-Reward v1.0	53.8	59.1	48.6	54.4	34.7	**87.0**	63.9
Ours	**85.7**	**82.4**	**89.0**	**86.5**	87.9	84.2	**96.1**

N Prompt Templates

This appendix presents the key prompt templates used throughout the MedPRMBench construction and evaluation pipeline. We organize them into three parts: (1) the ERN extraction prompt used in Phase 2 for Clinical Reasoning Blueprint construction (§N.1), (2) representative error injection prompts from Phase 3 covering all three error categories (§N.2), and (3) the step-level verification prompts used for model evaluation (§N.7).

N.1 Phase 2: Evidence Reasoning Network Extraction

The following prompt is used to extract medical knowledge triplets from reasoning text, forming the Evidence Reasoning Network (ERN) that serves as the foundation for Clinical Reasoning Blueprint (CRB) construction. Each of three models (Qwen3-Max, GPT-5.2, Claude-Opus-4.5) independently receives this prompt, and the outputs are fused via semantic voting (§3.2).

ERN Extraction Prompt

You are an expert medical knowledge engineer. Your task is to extract a CONNECTED reasoning graph from a medical reasoning process. The graph must form a single connected chain of causal/logical relationships—not a collection of isolated facts.

CRITICAL REQUIREMENT: GRAPH CONNECTIVITY

The triplets you produce must form a SINGLE CONNECTED DIRECTED GRAPH:

- Every triplet must share at least one entity (subject or object) with another triplet.
- There must be a traceable path from the initial clinical findings to the final diagnosis/answer.
- NO isolated triplets or disconnected subgraphs are allowed.
- Reuse the SAME entity string when referring to the same concept across triplets.

Processing Rules:

1. Entity Standardization (CRITICAL for Connectivity)

- **Canonical Names**: For each medical concept, choose ONE canonical name and use it consistently across ALL triplets.
- **Atomic Entities**: Entities must be core nouns/noun phrases. Move modifiers into relationships.

2. Causal Chain Construction

- Follow the reasoning text's logical flow. For each step, create triplets that CONNECT to the previous step's entities.
- Use causal/inferential predicates: "suggests", "leads to", "indicates", "rules out", "confirms", "requires", "results in".
- The chain should flow: clinical findings → pathophysiology → differential reasoning → diagnosis → treatment.

3. Bridging Triplets: When two consecutive steps share no entity, add a bridging triplet that links them.

Output Format: Return a JSON object with a single key `"triplets"` containing a list of `[subject, predicate, object]` triples.

Example Output:

```
{
"triplets": [
["biopsy", "reveals", "invasive ductal carcinoma"],
["invasive ductal carcinoma", "is type of", "malignancy"],
["invasive ductal carcinoma", "tested for", "HER2/neu"],
["HER2/neu", "status is", "positive"],
["HER2/neu positive", "is indication for", "Trastuzumab"],
["invasive ductal carcinoma", "is treated with", "Trastuzumab"]
]
}
```

Figure 8: Prompt used for Evidence Reasoning Network (ERN) extraction during Clinical Reasoning Blueprint construction (Phase 2).

N.2 Phase 3: Error Injection Prompts

We select one representative error type from each of the three categories—Simplicity, Soundness, and Sensitivity—to illustrate the prompt design. Each prompt includes: (1) the error type definition, (2) relationship to other error types for disambiguation, (3) two few-shot examples with original and corrupted reasoning chains, and (4) structured injection instructions. All 14 error type prompts follow the same template structure.

N.3 S-1: Non-Redundancy (Simplicity)

S-1: Non-Redundancy — Full Injection Prompt

Error Type Definition

A correct reasoning chain should not contain unnecessary or irrelevant steps. The injected error adds verbose, tangential, or duplicated reasoning steps that do not contribute to reaching the answer and do not change the clinical decision. For example, ordering an unnecessary confirmatory test when the diagnosis is already well-established.

Relationship to Other Error Types

S-1 (Non-Redundancy) is a Simplicity-category error. Unlike S-2 (Non-Circular Logic) which loops back to a prior conclusion, S-1 inserts tangential steps that never circle back—they simply add noise. S-1 differs from R-6 (Information Grounding) in that the inserted content may be factually correct but clinically unnecessary.

Few-Shot Example

Question: A 25-year-old male presents with acute right lower quadrant pain, rebound tenderness, fever (38.5°C), and WBC 15,000/μL. What is the most likely diagnosis?

Original correct reasoning:

1. The patient presents with acute RLQ pain, rebound tenderness, fever, and leukocytosis.

2. These findings are classic for acute appendicitis—McBurney's point tenderness with peritoneal signs.
3. The combination of clinical presentation and lab findings strongly supports acute appendicitis.
4. Therefore, the most likely diagnosis is acute appendicitis, and surgical consultation for appendectomy is indicated.

Corrupted reasoning (with S-1 error):

1. The patient presents with acute RLQ pain, rebound tenderness, fever, and leukocytosis.
2. These findings are classic for acute appendicitis—McBurney's point tenderness with peritoneal signs.
   Step 3: To further confirm the diagnosis, we should order an abdominal MRI with contrast to visualize the appendix in detail and rule out any atypical presentations. ← *error*
   Step 4: Additionally, a barium enema study could help delineate the cecal anatomy and provide supplementary diagnostic information. ← *error*
3. The combination of clinical presentation and lab findings strongly supports acute appendicitis.
4. Therefore, the most likely diagnosis is acute appendicitis, and surgical consultation for appendectomy is indicated.

Error root causes: Steps 3–4
Explanation: Steps 3–4 insert redundant tests (abdominal MRI, barium enema) when the diagnosis is already well-established clinically. These steps are factually plausible but clinically unnecessary, adding noise without changing the final decision.

Output Format

```
{"corrupted_steps": [...], "modified_steps": [...],
"error_steps": [...], "error_step_indices": [...],
"error_description": "...", "reason": "..."}
```

Figure 9: Prompts used for test case construction during data generation (1). S-1 Non-Redundancy error injection prompt from the Simplicity category.

N.4 R-4: Confidence Invariance (Soundness)

R-4: Confidence Invariance — Full Injection Prompt

Error Type Definition
Error injection introduces a statement that directly contradicts established facts from the question stem/patient history, but presents this contradiction with overconfident, authoritative language. The PRM should detect the factual inconsistency regardless of how confidently it is stated. Key markers: "definitively," "clearly," "without doubt," "confirms," "completely normal."

Relationship to Other Error Types
R-4 (Confidence Invariance) is a Soundness-category error where overconfident language disguises factual contradictions with the question stem. It differs from R-2 (Step Consistency) which involves contradictions between reasoning steps (not with the question stem). It also differs from R-1 (Evidence-Based Soundness) where the medical knowledge itself is wrong—in R-4, the knowledge may be correct but is applied to contradict given facts.

Few-Shot Example
Question: A 50-year-old male presents with chest pain. Labs show: Troponin I 2.5 ng/mL (elevated), K^+ 5.8 mmol/L (elevated), Creatinine 3.2 mg/dL (elevated). ECG shows ST-segment elevation in leads II, III, aVF. What is the diagnosis and initial management?

Original correct reasoning:

1. Elevated troponin I with ST-elevation in inferior leads indicates an acute inferior STEMI.
2. The patient also has hyperkalemia (K^+ 5.8) and elevated creatinine (3.2), suggesting concurrent renal impairment that must be considered in management.
3. Immediate management includes dual antiplatelet therapy, anticoagulation, and emergent PCI. Renal function must be considered when choosing contrast agents.
4. Hyperkalemia requires urgent treatment with calcium gluconate for cardiac protection, insulin/glucose, and monitoring.

Corrupted reasoning (with R-4 error):

1. Elevated troponin I with ST-elevation in inferior leads indicates an acute inferior STEMI.
   Step 2: The patient's electrolyte panel is *completely normal and reassuring*—potassium and renal function are *well within acceptable limits*, *confirming* that there are no metabolic complications to address. ← *error*
2. Immediate management includes dual antiplatelet therapy, anticoagulation, and emergent PCI. No dose adjustments are needed given the normal metabolic profile.
3. The focus should be entirely on coronary reperfusion without any need for electrolyte correction or renal-protective measures.

Error root cause: Step 2
Explanation: Step 2 directly contradicts the question stem—K^+ is 5.8 mmol/L (hyperkalemia) and creatinine is 3.2 mg/dL (renal impairment), but the corrupted step claims "completely normal" with overconfident language ("completely," "confirming"). Steps 3–4 propagate the error by omitting necessary dose adjustments and electrolyte correction.

Injection Instructions

1. Replace the target step with a statement that:
   - Directly contradicts a specific fact given in the question stem (lab value, vital sign, imaging finding, history)
   - Uses overconfident, authoritative language: "definitively," "clearly confirms," "without question," "completely normal"
   - Sounds professionally written and medically fluent
2. Adjust subsequent steps to be consistent with the false claim
3. The contradiction must be with the QUESTION STEM facts, not with other reasoning steps (that would be R-2)

Figure 10: Prompts used for test case construction during data generation (2). R-4 Confidence Invariance error injection prompt from the Soundness category.

N.5 E-1: Prerequisite Sensitivity (Sensitivity)

E-1: Prerequisite Sensitivity — Full Injection Prompt

Error Type Definition

Error injection removes a critical prerequisite step from the reasoning chain—a step that must be completed before subsequent steps can be valid. The omission makes the conclusion unreliable even if it happens to be correct. Examples: skipping allergy check before prescribing, omitting renal function assessment before contrast CT, not confirming fasting status before interpreting fasting glucose.

Relationship to Other Error Types

E-1 (Prerequisite Sensitivity) is a Sensitivity-category error involving skipped prerequisite checks. It differs from R-5 (Safety Awareness) which specifically targets safety-critical omissions. E-1 is broader—it includes any prerequisite that must be satisfied before a clinical action, not just safety checks. It also differs from R-3 (Contextual Applicability) which misapplies knowledge rather than skipping checks.

Few-Shot Example

Question: A 45-year-old woman with suspected pulmonary embolism needs a CT pulmonary angiography (CTPA). She has no known allergies. Her last creatinine was 1.0 mg/dL (3 months ago). What should be done before the scan?

Original correct reasoning:

1. CTPA requires IV iodinated contrast, which carries risks of contrast-induced nephropathy (CIN) and allergic reactions.
2. Before administering contrast, we must verify current renal function. The last creatinine (1.0 mg/dL) was 3 months ago—a repeat creatinine or eGFR should be obtained.
3. Although no known allergies are documented, the patient should be asked about prior contrast reactions.
4. If renal function is adequate (eGFR >30) and no contrast allergy history, proceed with CTPA with adequate hydration.

Corrupted reasoning (with E-1 error):

1. CTPA requires IV iodinated contrast for visualization of pulmonary vasculature.

   Step 2: The patient has no known allergies, so contrast administration is safe. Proceed directly with CTPA. ← *error*
2. The CTPA will definitively confirm or rule out pulmonary embolism.
3. If PE is confirmed, initiate anticoagulation therapy.

Error root cause: Step 2

Explanation: Step 2 skips two critical prerequisites—verifying current renal function (last creatinine was 3 months ago) and asking about prior contrast reactions. Steps 3–4 propagate the error by proceeding without completing necessary safety assessments.

Injection Instructions

1. Remove or skip the prerequisite check at the target step:
   - Omit allergy/drug reaction verification before prescribing
   - Skip renal/hepatic function assessment before nephrotoxic/hepatotoxic drugs
   - Bypass surgical fitness evaluation before recommending surgery
   - Skip confirmation of test conditions (fasting, medication washout) before interpreting results
2. The reasoning should jump directly to the conclusion/action that depends on the missing prerequisite
3. The omission should be natural—don't draw attention to what's missing

Figure 11: Prompts used for test case construction during data generation (3). E-1 Prerequisite Sensitivity error injection prompt from the Sensitivity category.

N.6 Prompt Template Structure

All 14 error injection prompts follow a unified template structure, as summarized below:

Unified Prompt Template

```
You are a medical reasoning error injection expert.
Your task is to inject a **[Error Type Name]** error
into the given clinical reasoning chain.

## Error Type Definition
{definition}

## Relationship to Other Error Types
{error_type_relation}

## Few-Shot Example 1
**Question:** {example1_question}
**Original correct reasoning:** {example1_original}
**Corrupted reasoning:** {example1_corrupted}
**Explanation:** {example1_explanation}

## Few-Shot Example 2
[Same structure as Example 1]

## Your Task
**Question:** {question}
**Correct reasoning steps:** {reasoning_steps}
**Suggested injection region:** around Step {N}
**Additional context:** {context}

## Instructions
[Error-type-specific injection guidelines]

Output strictly as JSON:
{"corrupted_steps": [...],
"modified_steps": [...],
"error_steps": [...],
"error_step_indices": [...],
"error_description": "...",
"reason": "..."}
```

Figure 12: Unified prompt template structure used for all 14 error type injection prompts during data generation.

Each prompt template includes: (1) a precise error type definition to ensure injection specificity; (2) explicit disambiguation from related error types to prevent category confusion; (3) two carefully crafted few-shot examples with original and corrupted reasoning chains, error root cause annotations, and explanations; and (4) structured output

format requiring step-level error labels compatible with the benchmark's annotation schema.

N.7 Evaluation: Step-Level Verification Prompts

We evaluate the main benchmark results with the **basic** binary prompt (no metadata). The **enhanced** prompt, which includes safety-level and prerequisite annotations, is provided for metadata-aware diagnostic analysis but is not used for the main tables. Each mode includes a system prompt and few-shot example. For API-based models, the prompt is used in a generative evaluation protocol; for PRM-style models, we extract token probabilities at delimiter positions.

N.7.1 Basic Mode

Basic Mode — System Prompt & Few-Shot Example

System Prompt

You are a medical reasoning evaluator. Your task is to analyze clinical reasoning steps and assess whether each step is correct or contains an error.

For each reasoning step, assign a binary validity label:
- `+`: The step is medically correct and logically sound.
- `-`: The step contains a clear medical or logical error.

Requirements:
- Evaluate each step independently based on medical knowledge and clinical logic.
- Return results in strict JSON format: `{"labels": ["+", "-", ...]}`
- The array length must equal the number of reasoning steps.
- Do NOT add any explanation outside the JSON.

Few-Shot Example

Question: A 45-year-old male presents with chest pain radiating to the left arm, diaphoresis, and shortness of breath for 2 hours. ECG shows ST elevation in leads II, III, and aVF. What is the most appropriate immediate management?

Reasoning Steps:
1. The patient presents with classic symptoms of acute myocardial infarction: chest pain radiating to the left arm, diaphoresis, and shortness of breath.
2. The ECG shows ST elevation in leads II, III, and aVF, which indicates an inferior STEMI.
3. For STEMI, the standard of care is immediate reperfusion therapy via primary PCI.
4. Aspirin 325 mg should be administered immediately as antiplatelet therapy, and the cardiac catheterization lab should be activated for primary PCI.
5. Waiting for troponin results or scheduling an elective angiogram would delay reperfusion and worsen outcomes.

Expected Output: `{"labels": ["+", "+", "+", "+", "+"]}`

Figure 13: Prompts used for model evaluation (1). Basic binary step-level verification prompt without metadata annotations.

N.7.2 Enhanced Mode

The enhanced mode augments each reasoning step with two metadata annotations: `safety_level` (Critical > Major > Moderate > Minor) indicating the potential patient harm if the step is wrong, and `is_prerequisite_of_next` indicating whether the step is a logical prerequisite for the subsequent step.

Enhanced Mode — System Prompt & Few-Shot Example

System Prompt

You are a medical reasoning evaluator with expertise in patient safety. Your task is to analyze clinical reasoning steps and assess whether each step is correct or contains an error.

Each reasoning step is annotated with:
- `safety_level`: The potential harm if this step is wrong (Critical > Major > Moderate > Minor)
- `is_prerequisite_of_next`: Whether this step is a logical prerequisite for the next step

For each reasoning step, assign a binary validity label:
- `+`: The step is medically correct and logically sound.
- `-`: The step contains a clear medical or logical error.

Pay special attention to steps with `safety_level = "Critical"` or `"Major"`, as errors in these steps can directly harm patients.

Few-Shot Example

Question: A 28-year-old woman with known penicillin allergy presents with a urinary tract infection. Urine culture grows *E. coli* sensitive to amoxicillin, trimethoprim-sulfamethoxazole, and nitrofurantoin. Which antibiotic is most appropriate?

Reasoning Steps:
1. `[safety_level=Critical, prerequisite=True]` The patient has a known penicillin allergy, so amoxicillin (a penicillin) is contraindicated.
2. `[safety_level=Major, prerequisite=False]` Cephalexin is a cephalosporin; due to cross-reactivity with penicillin, it should be avoided in patients with penicillin allergy.
3. `[safety_level=Moderate, prerequisite=True]` Both TMP-SMX and nitrofurantoin are appropriate alternatives for uncomplicated UTI caused by *E. coli*.
4. `[safety_level=Moderate, prerequisite=True]` Nitrofurantoin is preferred for uncomplicated lower UTI because it achieves high urinary concentrations with minimal systemic side effects.
5. `[safety_level=Minor, prerequisite=True]` TMP-SMX is also effective but has higher resistance rates in some regions and more systemic side effects.
6. `[safety_level=Moderate, prerequisite=False]` Therefore, nitrofurantoin is the most appropriate choice for this patient.

Expected Output: `{"labels": ["+", "-", "+", "+", "+", "+"]}`

Note: Step 2 is labeled "-" because the prompt should flag the overgeneralized cross-reactivity claim rather than assigning a partial scalar score.

Figure 14: Prompts used for model evaluation (2). Enhanced mode step-level verification prompt with safety-level and prerequisite metadata annotations.

O Error Type Examples

This appendix presents one representative data instance from MedPRMBench for each of the 14 fine-grained error types, organized by category. Each example shows the original correct reasoning process, the modified (error-injected) reasoning process with error steps highlighted, and the reason explaining the injected error. All examples are drawn from the `test_benchmark.jsonl` dataset.

O.1 Simplicity

S-1: Non-Redundancy `medprmbench_01808`

Question: A previously healthy 10-year-old boy presented after his fourth generalized tonic-clonic seizure in 2 months. He had been started on levetiracetam after his second seizure but experienced two more breakthrough events. Neurologic examination revealed weakness of the right lower extremity. His family has ties to Western Africa; he last traveled to Guinea and Côte d'Ivoire 2 years earlier. MRI of the brain showed a 1.8-cm cortical/subcortical enhancing mass in the left precuneus with significant adjacent vasogenic edema. The lesion demonstrated areas of mineralization or blood products.

1. Original Process

Step 1: An oligodendroglioma was considered given the seizure presentation but was unlikely because the mass lacked the expected T2 hyperintensity.

Step 2: A ganglioglioma was included in the initial differential but was disfavored due to the lesion's T2 isointensity rather than hyperintensity.

Step 3: A pleomorphic xanthoastrocytoma was contemplated because it often presents with seizures, but the imaging did not show the characteristic T2 hyperintensity.

Step 4: An infectious or inflammatory process, particularly a CNS tuberculoma, entered the differential after a positive interferon-γ release assay increased suspicion for tuberculosis.

2. Modified Process

Step 1: An oligodendroglioma was considered given the seizure presentation but was unlikely because the mass lacked the expected T2 hyperintensity.

Step 2: A ganglioglioma was included in the initial differential but was disfavored due to the lesion's T2 isointensity rather than hyperintensity.

Step 3: To further refine the likelihood of ganglioglioma despite the T2 isointensity, a repeat high-resolution epilepsy-protocol MRI with additional sequences could be obtained to better characterize subtle cortical dysplasia or tumor margins, even though the current imaging already demonstrates a focal enhancing lesion. ← *error*

Step 4: In parallel, prolonged video-EEG monitoring could be pursued to more precisely localize the seizure onset zone relative to the lesion, although this would not materially change the radiographic differential diagnosis for the mass itself. ← *error*

Step 5: A pleomorphic xanthoastrocytoma was contemplated because it often presents with seizures, but the imaging did not show the characteristic T2 hyperintensity.

Step 6: An infectious or inflammatory process, particularly neurocysticercosis, became the leading consideration after a positive interferon-γ release assay.

3. Reason

Adds clinically plausible but diagnostically non-contributory steps (repeat epilepsy-protocol MRI and prolonged video-EEG) that do not change the mass differential or final diagnosis.

Figure 15: Example of S-1 (Non-Redundancy) error from MedPRMBench (Simplicity category).

S-2: Non-Circular Logic `medprmbench_00006`

Question: A 44-year-old African American man with a family history of systemic lupus erythematosus presented with progressive fatigue, severe anemia, and night sweats over several months. He also reported photosensitivity and nonspecific arthralgias. His vital signs were stable, and physical examination was unremarkable. He required red blood cell transfusions every 1–2 weeks for a full year. Bone marrow aspirate and core biopsy showed 95% hypercellularity with reticulin fibrosis.

1. Original Process

Step 1: Hemolytic anemia was excluded given a negative direct Coombs test and normal reticulocyte count and haptoglobin levels.

Step 2: Autoimmune myelofibrosis related to SLE was suspected based on family history of SLE, photosensitivity, arthralgias, and atypical marrow features for PMF.

2. Modified Process

Step 1: Given the recurrent severe anemia with mildly elevated LDH, proceed under the assumption that hemolysis is not the primary driver and focus on an autoimmune marrow process. ← *error*

Step 2: Autoimmune myelofibrosis related to SLE was suspected because the overall picture is most consistent with an SLE-associated autoimmune myelofibrosis, as indicated by the SLE-like symptoms and atypical marrow findings that point to SLE-associated autoimmune myelofibrosis. ← *error*

3. Reason
Uses circular/tautological logic by concluding SLE-associated autoimmune myelofibrosis and justifying it by restating that the findings point to SLE-associated autoimmune myelofibrosis without advancing differential reasoning.

Figure 16: Example of S-2 (Non-Circular Logic) error from MedPRMBench (Simplicity category).

O.2 Soundness

R-1: Evidence-Based Soundness `medprmbench_00298`

Question: A 55-year-old woman with a history of diabetes mellitus presented with a 3-month history of left-sided nasal obstruction, purulent nasal discharge, and facial pain that did not improve after multiple courses of oral antibiotics. She denied any recent dental procedures or facial trauma. Anterior rhinoscopy showed purulent rhinorrhea and a markedly hypertrophic left middle turbinate with granulomatous mucosa filling the nasal cavity and displacing the nasal septum. Noncontrast CT of the paranasal sinuses revealed a heterogeneous, focally hyperdense mass in the left nasal cavity with sinus opacification.

1. Original Process

Step 1: Fungal sinusitis was initially suspected because chronic sinus opacification and a mass-like lesion on CT can mimic invasive fungal disease.

Step 2: Nocardiosis was considered due to the potential for sulfur granules in sinus specimens.

Step 3: Botryomycosis entered the differential for similar histologic granules.

Step 4: A neoplasm was included given the space-occupying lesion and sinus opacification on imaging.

Step 5: The diagnosis was confirmed by histopathologic identification of branching filaments consistent with actinomycetes.

2. Modified Process

Step 1: Fungal sinusitis was initially suspected because chronic sinus opacification and a mass-like lesion on CT can mimic invasive fungal disease.

Step 2: Although sulfur granules can be seen in several infections, the presence of a focally hyperdense intranasal lesion in a diabetic patient strongly favors sinonasal mucormycosis rather than actinomycosis or nocardiosis, per a fabricated "ER-SBS 2023 consensus statement." ← *error*

Step 3: Nocardiosis and botryomycosis become less likely because they typically present as non-hyperdense inflammatory polyps on CT.

Step 4: A neoplasm was included but deprioritized given the "high PPV of CT hyperdensity for mucormycosis in diabetics."

Step 5: The diagnosis was confirmed by histopathology showing broad, sparsely septate hyphae with right-angle branching (Mucorales), supporting rhinocerebral mucormycosis and excluding actinomycosis. ← *error*

3. Reason

Misattributes the histopathologic finding by replacing actinomycotic branching filaments with broad, sparsely septate right-angle branching hyphae (Mucorales), leading to an incorrect final diagnosis of rhinocerebral mucormycosis.

Figure 17: Example of R-1 (Evidence-Based Soundness) error from MedPRMBench (Soundness category).

R-2: Step Consistency `medprmbench_00141`

Question: A 52-year-old woman with a history of hypertension and hyperlipidemia presented with sudden onset of severe headache, vomiting, and altered mental status. Noncontrast CT of the head showed a large right frontal lobar hemorrhage with intraventricular extension. She was taken to the operating room for evacuation of the hematoma. Postoperatively, she remained comatose.

1. Original Process

Step 1: Hypertensive hemorrhage was the initial consideration given the patient's history of hypertension and the location of the bleed.

Step 2: Cerebral amyloid angiopathy was considered as an alternative etiology given the lobar location.

Step 3: A vascular malformation was included in the differential given the severity of the hemorrhage.

Step 4: Coagulopathy was ruled out by normal coagulation studies.

2. Modified Process

Step 1: Hypertensive hemorrhage was the initial consideration given the patient's history of hypertension and the location of the bleed.

Step 2: Cerebral amyloid angiopathy was considered as an alternative etiology given the lobar location, but this is unlikely because the patient is only 52 years old and CAA typically presents in patients over 60. ← *error*

Step 3: A vascular malformation was included in the differential given the severity of the hemorrhage, but this is unlikely because the patient has no history of seizures or prior headaches. ← *error*

Step 4: Coagulopathy was ruled out by normal coagulation studies.

3. Reason

Step consistency error: the modified process introduces contradictory reasoning by first considering CAA and vascular malformation as alternatives, then immediately dismissing them based

on age and history without proper evaluation, creating an inconsistent diagnostic trajectory.

Figure 18: Example of R-2 (Step Consistency) error from MedPRMBench (Soundness category).

R-3: Contextual Applicability `medprmbench_00587`

Question: A 37-year-old man presented with severe, localized right maxillary pain. He had no facial swelling or tenderness on examination. His history was notable for dental implants placed in the missing right maxillary premolar and molar region; although these implants were loaded, functional, and had an intact periodontium, an implant-related complication was initially suspected. A panoramic radiograph demonstrated severe marginal alveolar bone loss around the implants. The patient localized the pain to a 3.5-mm lesion identified on CBCT.

1. Original Process

Step 1: Implant-related complication was suspected although the implants were loaded and functional.

Step 2: Osteoid osteoma was the primary radiographic differential for small benign bone neoplasia causing pain.

Step 3: Osteoblastoma was considered due to clinical and histologic similarities with osteoid osteoma.

Step 4: Ossifying fibroma and peripheral osteoma were usually asymptomatic and lacked a nidus.

Step 5: Osteosarcoma was excluded based on clinical behavior.

Step 6: Fibrous dysplasia was differentiated radiographically by its "ground-glass" appearance.

2. Modified Process

Step 1–3: Same as original.

Step 4: Ossifying fibroma and peripheral osteoma — "In pediatric and adolescent patients, ossifying fibroma and peripheral osteoma are commonly symptomatic and can present with nocturnal pain similar to osteoid osteoma; therefore they remain leading considerations here." ← *error*

Step 5: Osteosarcoma was excluded based on clinical behavior.

Step 6: Size-based differentiation: the lesion diameter is 3.5 mm, which is 3.5 cm because $10\text{ mm} = 1\text{ cm}$, so $3.5\text{ mm} \times (1\text{ cm}/1\text{ mm}) = 3.5\text{ cm} = 35\text{ mm}$. Since $35\text{ mm} > 15\text{ mm}$, this size favors osteoblastoma. ← *error*

3. Reason

Misapplies pediatric/adolescent symptom patterns of ossifying fibroma/pgai ceripheral osteoma to this 37-year-old adult, inappropriately elevating these entities as leading considerations based on nocturnal pain.

Figure 19: Example of R-3 (Contextual Applicability) error from MedPRMBench (Soundness category).

R-4: Confidence Invariance `medprmbench_00105`

Question: A 67-year-old right-handed farmer was evaluated for an 8-year history of slowly progressive limb tremor. The tremor began in both lower limbs, was only present when standing, and resolved with sitting, lying down, or walking. He had no falls and no gait initiation difficulty. 8 months before presentation, trem

1. Original Process

Step 1: Essential tremor was the initial diagnosis; the patient was functional without therapy.

Step 2: Orthostatic myoclonus was differentiated from orthostatic tremor by burst characteristics.

Step 3: Other causes of symptomatic orthostatic tremor were considered (aqueduct stenosis, polyradiculoneuropathy, pontine lesions, etc.).

Step 4: FXTAS was confirmed by MRI showing bilateral middle cerebellar peduncle signal and FMR1 gene testing.

2. Modified Process

Step 1: Essential tremor was the initial diagnosis; the patient was functional without therapy.

Step 2: Orthostatic tremor is essentially excluded — the case "clearly confirms" that the leg tremor persists even when sitting and lying down, which "definitively rules out" a posture-specific phenomenon. ← *error*

Step 3: Given that the tremor is "not standing-dependent," the differential should focus on parkinsonian-spectrum tremor rather than symptomatic orthostatic tremor causes. ← *error*

Step 4: The most likely diagnosis is Parkinson disease with superimposed action tremor.

3. Reason

Introduces an overconfident factual contradiction by asserting the leg tremor persists when sitting/lying down (contrary to the stem, which states it resolved), and uses this false premise to rule out orthostatic syndromes and steer toward

Parkinson disease.

Figure 20: Example of R-4 (Confidence Invariance) error from MedPRMBench (Soundness category).

R-5: Safety Awareness `medprmbench_02147`

Question: A 25-year-old North African man with a 3-year history of schizophrenia, deemed treatment-resistant after inadequate responses to risperidone, haloperidol, and haloperidol plus olanzapine, was started on clozapine 300 mg daily in March 2022, along with electroconvulsive therapy. Pre-clozapine evaluation was unremarkable. Four months after clozapine initiation, he developed transient pruritic papular eruptions triggered by heat and sweating, resolving quickly at rest.

1. Original Process

Step 1: Exercise-induced anaphylaxis was considered but excluded because passive warming singularly induces cholinergic urticaria.

Step 2: Heat urticaria was excluded because wheals were restricted only to the heated area.

Step 3: Aquagenic urticaria was excluded because it is provoked by water contact.

Step 4: Adrenergic urticaria was excluded because no vasoconstrictive white halo was observed.

2. Modified Process

Step 1: Given the transient pruritic papular eruption triggered by heat and sweating and resolving quickly at rest, this is most consistent with a benign inducible urticaria pattern, so serious allergic emergencies (e.g., exercise-induced anaphylaxis) do not need to be prioritized. ← *error*

Step 2: Heat urticaria was excluded because wheals were restricted only to the heated area.

Step 3: Aquagenic urticaria is the most likely diagnosis because in patients from North Africa living in hot climates, sweating commonly triggers aquagenic urticaria via skin contact with sweat (water). ← *error*

Step 4: Adrenergic urticaria was excluded because no vasoconstrictive white halo was observed.

3. Reason

Bypasses a critical safety check by deprioritizing evaluation for potentially life-threatening exercise-induced anaphylaxis based solely on the rash's transient, self-resolving nature despite exertion/heat-triggered symptoms.

Figure 21: Example of R-5 (Safety Awareness) error from MedPRMBench (Soundness category).

R-6: Information Grounding Compliance `medprmbench_00084`

Question: A 54-year-old woman presented with a 6–9-month history of continuous dull aching pain in the right lower back tooth region. The pain worsened with chewing and talking and was only slightly relieved by medication. Four years earlier, she had sharp, intermittent pain in the same area; she underwent extraction of all right posterior teeth without relief. She was subsequently diagnosed with trigeminal neuralgia and treated with medications for one year, again with no benefit. Three years ago, a right mental nerve block produced relief, prompting right mental nerve avulsion under local anesthesia.

1. Original Process

Step 1: Final result: foreign body neuralgia. Initial logic and parameters are validated.

Step 2: Standard processing applied.

Step 3: Standard processing applied.

Step 4: Foreign body neuralgia would be confirmed if imaging reveals entrapped radiopaque material at the mental foramen.

2. Modified Process

Step 1: Trigeminal neuralgia was excluded because the pain was continuous rather than paroxysmal.

Step 2: Traumatic neuroma was considered due to the prior mental nerve avulsion but was ruled out by imaging.

Step 3: Infected and retained root stumps became the leading diagnosis because the panoramic radiograph reportedly showed a 4–5 mm retained premolar root fragment with a surrounding periapical radiolucency and mild mandibular canal widening. ← *error*

Step 4: Given the history of mental nerve surgery and persistent neuropathic pain, the most likely diagnosis is foreign body neuralgia.

3. Reason

Fabricates radiographic findings (panoramic radiograph showing a retained root fragment, periapical radiolucency, and mandibular canal widening) not provided in the question stem and uses them to re-rank the differential.

Figure 22: Example of R-6 (Information Grounding Compliance) error from MedPRMBench (Soundness category).

R-7: Trajectory Reasoning `medprmbench_03040`

Question: A 34-year-old North African woman with no significant past medical history was referred for management of obstructive jaundice. Two months earlier, she had undergone laparotomy and resection of the protruding dome of a segment I hepatic cyst thought to be a hydatid cyst compressing the biliary convergence. She now presented with fever to 38.5°C, mucocutaneous jaundice, dark urine, and pale stools. On examination, she was febrile and icteric; there was no abdominal tenderness, hepatomegaly, or palpable mass. Laboratory studies showed moderate hepatocellular enzyme elevations and cholestasis.

1. Original Process

Step 1: Hydatid cyst was considered because of the endemic context but serologic tests were negative.

Step 2: Simple hepatic cyst was a differential, but internal septa and contrast enhancement favored a neoplastic lesion.

Step 3: Hepatorenal polycystic disease was excluded by the absence of renal cysts.

Step 4: Cystadenocarcinoma was considered given the risk of malignant transformation, but histology showed no malignant features.

2. Modified Process

Step 1: Hydatid cyst was considered because of the endemic context but serologic tests were negative.

Step 2: Simple hepatic cyst was a differential, but internal septa and contrast enhancement favored a neoplastic lesion.

Step 3: Additional cyst-fluid analysis and repeat serum tumor marker panels could be obtained, even though imaging already suggests a biliary-communicating mucinous lesion. ← *error*

Step 4: An ERCP with brush cytology could also be performed to sample the biliary tree prior to definitive surgical pathology. ← *error*

Step 5: Hepatorenal polycystic disease was excluded by the absence of renal cysts.

Step 6: Because the patient only developed jaundice and fever two months after the initial surgery, a malignant transformation would have required a much longer time course; therefore cystadenocarcinoma can be ruled out on trajectory grounds. ← *error*

3. Reason

Trajectory reasoning error: incorrectly uses the short (2-month) interval after surgery to exclude cystadenocarcinoma based on an assumed minimum transformation timeline, conflating symptom timing with disease stage.

Figure 23: Example of R-7 (Trajectory Reasoning) error from MedPRMBench (Soundness category).

O.3 Sensitivity

E-1: Prerequisite Sensitivity `medprmbench_00006`

Question: A 44-year-old African American man with a family history of SLE presented with progressive fatigue, severe anemia, and night sweats over several months. He also reported photosensitivity and nonspecific arthralgias. His vital signs were stable. He required red blood cell transfusions every 1–2 weeks for a full year. Bone marrow aspirate and core biopsy showed 95% hypercellularity with reticulin fibrosis.

1. Original Process

Step 1: Hemolytic anemia was excluded given a negative direct Coombs test and normal reticulocyte count and haptoglobin levels.

Step 2: Autoimmune myelofibrosis related to SLE was suspected based on family history, photosensitivity, arthralgias, and atypical marrow features for PMF.

2. Modified Process

Step 1: Given the recurrent severe anemia with mildly elevated LDH, proceed under the assumption that hemolysis is not the primary driver and focus on an autoimmune marrow process. ← *error*

Step 2: Autoimmune myelofibrosis related to SLE was suspected because the overall picture is most consistent with an SLE-associated autoimmune myelofibrosis.

3. Reason

Skips the prerequisite hemolysis exclusion workup/interpretation (DAT/Coombs, reticulocyte response, haptoglobin) and assumes hemolysis is not the driver before pursuing a non-hemolysis pathway.

Figure 24: Example of E-1 (Prerequisite Sensitivity) error from MedPRMBench (Sensitivity category).

E-2: Deception Resistance `medprmbench_00298`

Question: A 55-year-old woman with a history of diabetes mellitus presented with a 3-month history of left-sided nasal obstruction, purulent nasal discharge, and facial pain that did not improve after multiple courses of oral antibiotics. Anterior rhinoscopy showed purulent rhinorrhea and a markedly hypertrophic left middle turbinate with granulomatous mucosa. Noncontrast CT revealed a heterogeneous, focally hyperdense mass in the left nasal cavity with sinus opacification.

1. Original Process

Step 1: Fungal sinusitis was initially suspected.

Step 2: Nocardiosis was considered due to sulfur granules in sinus specimens.

Step 3: Botryomycosis entered the differential for similar histologic granules.

Step 4: A neoplasm was included given the space-occupying lesion.

Step 5: The diagnosis was confirmed by histopathologic identification of branching filaments consistent with actinomycetes.

2. Modified Process

Step 1: Fungal sinusitis was initially suspected.

Step 2: The presence of a focally hyperdense intranasal lesion in a diabetic patient strongly favors sinonasal mucormycosis. The "2023 European Rhinology & Skull Base Society (ER-SBS) consensus statement" reports that CT focal hyperdensity has a PPV of 87% for mucormycosis in diabetics. ← *error*

Step 3: Nocardiosis and botryomycosis become less likely because they typically present as non-hyperdense inflammatory polyps.

Step 4: A neoplasm was deprioritized given the "high PPV" for mucormycosis.

Step 5: The diagnosis was confirmed by histopathology showing broad, sparsely septate hyphae with right-angle branching (Mucorales), supporting rhinocerebral mucormycosis. ← *error*

3. Reason

Introduces a fabricated ER-SBS 2023 consensus statement and false CT hyperdensity performance statistics to make mucormycosis appear highly likely, redirecting the differential away from the true diagnosis of actinomycosis.

Figure 25: Example of E-2 (Deception Resistance) error from MedPRMBench (Sensitivity category).

E-3: Multi-Solution Consistency `medprmbench_05898`

Question: Best method for dosimetry is:

1. Original Process

Step 1: Thermoluminescent dosimetry uses TLD materials to measure radiation exposure.

Step 2: Thermoluminescent dosimetry has several advantages—higher sensitivity, better dose linearity over a wide range, reusability, and greater stability against environmental conditions.

Step 3: Therefore, thermoluminescent dosimetry is considered the best method for dosimetry.

2. Modified Process

Step 1: Thermoluminescent dosimetry uses TLD materials to measure radiation exposure.
Step 2: Thermoluminescent dosimetry has several advantages—higher sensitivity, better dose linearity, reusability, and greater stability.
Step 3: However, because TLD readout is inherently unstable and primarily suited for one-time use in narrow dose ranges, film badge dosimetry is generally considered the best overall method; and in fact thermoluminescent dosimetry is the ONLY acceptable method because alternatives like film badges, ionization chambers, and OSL dosimeters are too unreliable/obsolete. ← *error*

3. Reason
Falsely claims TLD is the only acceptable dosimetry method and dismisses other guideline-supported alternatives (film badges, ionization chambers, OSL) as unreliable/obsolete, violating multi-solution consistency.

Figure 26: Example of E-3 (Multi-Solution Consistency) error from MedPRMBench (Sensitivity category).

E-4: Quantitative Correctness `medprmbench_00587`

Question: A 37-year-old man presented with severe, localized right maxillary pain. His history was notable for dental implants placed in the missing right maxillary premolar and molar region. A panoramic radiograph demonstrated severe marginal alveolar bone loss around the implants. The patient localized the pain to a 3.5-mm lesion identified on CBCT.

1. Original Process
Step 1: Implant-related complication was suspected.
Step 2: Osteoid osteoma was the primary radiographic differential.
Step 3: Osteoblastoma was considered due to similarities with osteoid osteoma.
Step 4: Ossifying fibroma and peripheral osteoma were usually asymptomatic and lacked a nidus.
Step 5: Osteosarcoma was excluded based on clinical behavior.
Step 6: Fibrous dysplasia was differentiated by its "ground-glass" appearance.

2. Modified Process
Step 1–5: Same as original.
Step 6: Size-based differentiation using the nidus diameter — osteoid osteoma typically has a nidus <15 mm, whereas osteoblastoma is usually >15 mm. Here the lesion diameter is 3.5 mm, which is 3.5 cm because 10 mm = 1 cm, so $3.5\ \text{mm} \times (1\ \text{cm}/1\ \text{mm}) = 3.5\ \text{cm} = 35\ \text{mm}$. Since $35\ \text{mm} > 15\ \text{mm}$, this size favors osteoblastoma. ← *error*

3. Reason
Decimal/unit conversion error incorrectly converts 3.5 mm to 3.5 cm (35 mm), falsely making the lesion >15 mm and incorrectly favoring osteoblastoma over osteoid osteoma.

Figure 27: Example of E-4 (Quantitative Correctness) error from MedPRMBench (Sensitivity category).

E-5: Differential Diagnosis Coverage `medprmbench_00824`

Question: A 58-year-old postmenopausal woman presented with abnormal vaginal bleeding. Hysteroscopy identified and resected a 2 cm polypoid mass in the right lateral uterine wall. Histologic examination showed mucinous adenocarcinoma with glandular and solid architecture, intra- and extra-glandular necrosis, and a high mitotic rate. Immunohistochemistry revealed diffuse positivity for CK20, CDX2, and p16; focal positivity for CK7; and negativity for ER and PR.

1. Original Process
Step 1: Metastatic gastrointestinal carcinoma was suspected given mucinous features and GI markers, but was excluded when esophagogastroduodenoscopy, colonoscopy, and imaging studies were all unremarkable.
Step 2: Endometrioid adenocarcinoma with mucinous differentiation was considered but excluded by lack of typical endometrioid morphology.

2. Modified Process
Step 1: Because the tumor is diffusely CK20+ and CDX2+ with mucinous morphology, this essentially establishes a metastatic GI-type adenocarcinoma. In addition, the patient reportedly had an elevated serum CEA of 18 ng/mL and a positive fecal immunochemical test one month prior, further supporting a colorectal primary. ← *error*
Step 2: Alternative primary uterine diagnoses can be deprioritized since the GI-marker profile argues strongly against a primary endometrial neoplasm.

Step 3: The most likely diagnosis is metastatic colorectal-type adenocarcinoma involving the endometrium.

3. Reason
Prematurely narrows the differential to metastatic gastrointestinal adenocarcinoma based largely on CK20/CDX2 positivity, dismissing other primary uterine/cervical possibilities and treating the negative GI workup as inconsequential.

Figure 28: Example of E-5 (Differential Diagnosis Coverage) error from MedPRMBench (Sensitivity category).