PEDMED-BENCH: A MULTI-TURN BENCHMARK FOR LARGE LANGUAGE MODELS ON PEDIATRIC MEDICATION EFFECTIVENESS AND SAFETY

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating Medical Large Language Models (LLMs) in complex clinical scenarios has gained increasing attention. Yet most existing benchmarks emphasize general or adult-centered medicine, making them poorly aligned with pediatric care, where age-dependent physiology, longitudinal disease evolution, and medication safety require specialized reasoning. Specifically, current evaluation systems regarding pediatric care are still facing severe data scarcity challenges, thereby hindering the evaluation and development of pediatric Medical LLMs. Motivated by this, we introduce a novel benchmark (PedMed-Bench) to address the above challenges. Further, we conduct a systematic evaluation involving numerous representative LLMs on PedMed-Bench. The results suggest that contemporary LLMs still struggle to accurately predict adverse drug events (ADEs) and suspected drugs. Even though models equipped with reasoning capabilities demonstrate notable enhancements, their absolute performance remains constrained, with the highest ADE extraction F1 score reaching only 12.89%. To the best of our knowledge, PedMed-Bench is the first fine-grained, event-driven benchmark for real-world pediatric medication evaluation. The benchmark is released in both Chinese and English, with human-verified translation validated for cross-lingual equivalence. We hope it can serve as a helpful resource, contributing to the ongoing efforts in developing more reliable clinical AI tools for pediatric healthcare.

1 INTRODUCTION

Large language models (LLMs) have emerged as foundation models for diverse clinical tasks, including data integration, information synthesis, diagnostic support, and treatment planning (Moor et al., 2023; Thirunavukarasu et al., 2023; Van Veen et al., 2024; Shmatko et al., 2025). As shown in Table 1, early efforts primarily targeted medical question-answering benchmarks such as MedQA and PubMedQA (Singhal et al., 2023; 2025). To better capture the dynamic, iterative nature of real-world clinical reasoning (Bedi et al., 2025; Chen et al., 2025), recent work has introduced workflow-aligned benchmarks that incorporate progressive clinical stages and safety-critical interactions (McCoy et al., 2025; Zhou et al., 2025; Johri et al., 2025; Qiu et al., 2025; Arora et al., 2025). Despite these innovations, current benchmarks lack sufficient coverage of pediatric practice, a key and unique dimension of comprehensive clinical evaluation (Zhu et al., 2025a).

Compared with adult care, pediatric care involves more frequent polypharmacy, iterative dose titration, and heightened sensitivity to dosing due to age- and weight-dependent physiology Feudtner et al. (2012); Kearns et al. (2003). Ethical and practical constraints on pediatric trials limit clinical evidence, leading to widespread unlicensed or off-label prescribing (over 40% of prescriptions) and complicating pharmacovigilance Conroy et al. (2000). Consequently, pediatrics represents both a high-need domain for clinical decision support Kaushal et al. (2001) and a critical testbed for evaluating LLMs in realistic, safety-sensitive settings. However, existing benchmarks either lack pediatric cases or fail to capture the unique diagnostic patterns, therapeutic considerations, and safety challenges inherent to pediatric medication management. This gap hinders the development and assessment of LLMs that can effectively support pediatric care, where accurate medication evaluation is crucial for optimizing outcomes and minimizing risks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

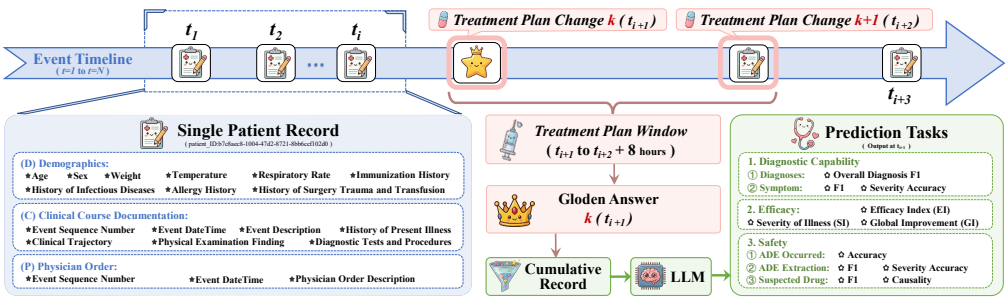


Figure 1: Overview of the PedMed-Bench framework for pediatric medication evaluation. The framework mirrors real-world clinical practice through expert-defined "Effective Observation Windows" that model residual drug effects.

To address these challenges, as shown in Figure 1, we introduce Pediatric Medication Benchmark (PedMed-Bench), the first benchmark for pediatric medication reasoning built from real-world inpatient Electronic Health Records (EHRs) reflecting clinical workflows. PedMed-Bench asks whether LLMs can make medication judgments that are temporally grounded, causally sensitive, and developmentally aware. It focuses on three dimensions: (i) Diagnostic Capability, examining LLMs' synthesis of complex pediatric evidence through the evolving trajectory of diagnoses, symptoms, and three-level severity. It measures the models' ability to progress from fragmented information retrieval to coherent diagnostic reasoning in cases with up to eight comorbidities; (ii) Therapeutic Efficacy, assessing temporal reasoning over extended clinical courses using standardized clinical scales and expert-defined Effective Observation Windows. This tests models' capacity to detect subtle temporal trends within extensive procedural records to predict the future impact of current treatments; and (iii) Medication Safety, probing causal inference under high-stakes conditions aligned with international pharmacovigilance standards. By incorporating counterfactual reasoning tasks, this component evaluates whether models can eliminate confounders and identify drugs responsible for adverse events Hazell & Shakir (2006).

To mitigate underreporting of safety outcomes, we extract EHR data from two tertiary hospitals and reconstruct Adverse Drug Events (ADEs) and Adverse Drug Reactions (ADRs) under expert supervision. To ensure contextual completeness, we integrate prescriptions, related conditions, and longitudinal trajectories into unified multi-view patient profiles. Each hospitalization is segmented into treatment episodes anchored to regimen changes, maintaining causal continuity between disease progression and medication adjustment for dynamic belief updating. Finally, we establish a clinically aligned evaluation framework with 11 granular indicators of diagnostic accuracy, therapeutic efficacy, and medication safety, validated by senior clinicians for evidence-based reliability.

We evaluate 14 representative LLMs across PedMed-Bench's multidimensional metrics, leading to three key takeaways: (1) *current LLMs remain limited in pediatric medication reasoning, especially for fine-grained safety attribution*; (2) *temporal perturbation and counterfactual medication interventions reveal that models are weakly grounded in chronological and medication-level causal evidence*; (3) *scaling, prompting, retrieval, memory, and reasoning-oriented inference yield task-dependent gains but do not reliably solve pediatric medication safety reasoning*.

Main contributions of this paper are as follows:

- We identify pediatric medication reasoning as a critical evaluation gap, where developmental physiology, dynamic treatment adjustment, and medication safety create challenges beyond adult-centered benchmarks.
- To the best of our knowledge, we construct PedMed-Bench, the first real-world inpatient EHR benchmark for pediatric medication effectiveness and safety, with expert-annotated treatment episodes and clinically grounded observation windows.
- We systematically evaluate 14 representative LLMs and show that current models struggle with fine-grained safety attribution, temporal grounding, and medication-level causal reasoning, informing safer clinical AI development.

To facilitate further research, all datasets and code are publicly available at `https://anonymous.4open.science/r/PedMed-Bench-252F/`.

Table 1: Comparison between PedMed-Bench and representative medical benchmarks or datasets. "Multi-dim." denotes multi-dimensional evaluation beyond binary or single-label assessment.

Benchmark / Dataset	Population	Drug Scenarios	Drug Safety	Drug Efficacy	Dynamic Data	Developmental Reasoning
MedQA Singhal et al. (2023)	General	✘	✘	✘	✘	✘
PubMedQA Singhal et al. (2025)	General	✘	✘	Moderate	✘	✘
MedR-Bench Qiu et al. (2025)	General	✘	✘	✘	✔	✘
HealthBench Arora et al. (2025)	General	✔	Low	Moderate	✘	✘
MedCalc-Bench Khandekar et al. (2024)	General	✔	Moderate	✘	✘	✘
MIMIC-IV Johnson et al. (2020)	Adult	✔	Low	Low	✔	✘
ER-REASON Mehandru et al. (2025)	ICU/Emergency	✔	Moderate	Moderate	✘	✘
PedMed-Bench	**Pediatric**	✔	**Multi-dim.**	**Multi-dim.**	✔	✔

2 Related Work

Medical LLM evaluation has shifted from static knowledge recall toward clinical reasoning, decision support, and safety-critical deployment (Bedi et al., 2025; Shmatko et al., 2025; Zhao et al., 2025). Early QA benchmarks such as MedQA (Singhal et al., 2023) and PubMedQA (Singhal et al., 2025) established standardized tests of biomedical knowledge, while recent benchmarks such as DiagnosisArena (Zhu et al., 2025b), MedR-Bench (Qiu et al., 2025), and MedAgentBench (Jiang et al., 2025) introduce more realistic diagnostic workflows, stepwise reasoning, and EHR-based interaction. Other resources, including HealthBench (Arora et al., 2025), MIMIC-IV (Johnson et al., 2020), ER-REASON (Mehandru et al., 2025), and MedCalc-Bench (Khandekar et al., 2024), further cover safety, dynamic clinical data, or medication-related scenarios.

Medication evaluation requires more than selecting an appropriate treatment at a single time point. Models must integrate longitudinal symptoms, regimen changes, delayed therapeutic effects, and potential adverse drug events, often under confounding from disease progression and polypharmacy. These challenges are amplified in pediatrics, where drug disposition, dosing, and adverse-event risk vary with developmental stage, age, and weight (Kearns et al., 2003); hospitalized children are frequently exposed to polypharmacy (Feudtner et al., 2012); and limited pediatric trial evidence contributes to widespread off-label or unlicensed prescribing (Conroy et al., 2000). Although medication errors and adverse drug events are a central safety concern in pediatric inpatient care (Kaushal et al., 2001), existing LLM benchmarks rarely evaluate drug efficacy, ADE/ADR attribution, severity grading, and medication-level causality together in dynamic pediatric trajectories.

As summarized in Table 1, existing benchmarks largely target general or adult populations and rarely combine pediatric medication safety, treatment efficacy, dynamic inpatient trajectories, and developmental reasoning. PedMed-Bench fills this gap by evaluating LLMs on pediatric inpatient medication trajectories, requiring models to reason over treatment adjustments, ADE/ADR attribution, therapeutic response, and age-dependent physiological variation.

3 Problem Formulation

We formulate the pediatric inpatient care process as a **sequential decision evaluation** task. Given a patient's longitudinal clinical trajectory $\mathcal{T} = \{x_1, \ldots, x_T\}$, we partition $\mathcal{T}$ into K treatment episodes, with $K < T$, indexed by $k \in \{1, \ldots, K\}$. For episode k, we define the model input as $\mathcal{X}_k = (\mathbf{C}_k, \mathcal{H}_{<k})$. Here, $\mathbf{C}_k$ denotes the patient state at the beginning of episode k, aggregating demographic and clinical variables such as age, gender, medical history, and physical examination findings. The term $\mathcal{H}_{<k}$ denotes the historical context prior to episode k, including information from earlier episodes: previous treatments $\mathbf{Tr}_i$ for $i < k$ (e.g., progress notes, physical findings, and treatment orders) and previous patient states $\mathbf{C}_i$ for $i < k$.

The objective is to predict a multi-dimensional evaluation target $\mathcal{Y}_k$ from $\mathcal{X}_k$. Ground-truth labels $\mathcal{Y}_k$ are constructed from outcomes observed within an "Effective Observation Window" $[t_k,\ t_{k+1} + 8\text{h}]$. Specifically, the target $\mathcal{Y}_k$ comes from three tasks: (1) **Symptomatology Diagnosis Effectiveness** predicts diagnosis ($\mathcal{Y}_{diag}$), symptom–severity pairs ($\mathcal{Y}_{sym}$) with ordinal levels $l \in \{0, 1, 2\}$; (2) **Therapeutic Efficacy Prediction** predicts a three-dimensional ordinal vector ($\mathcal{Y}_{eff}$) aligned with Clinical Global Impression (CGI) metrics, including Severity of Illness (SI), Global Improvement

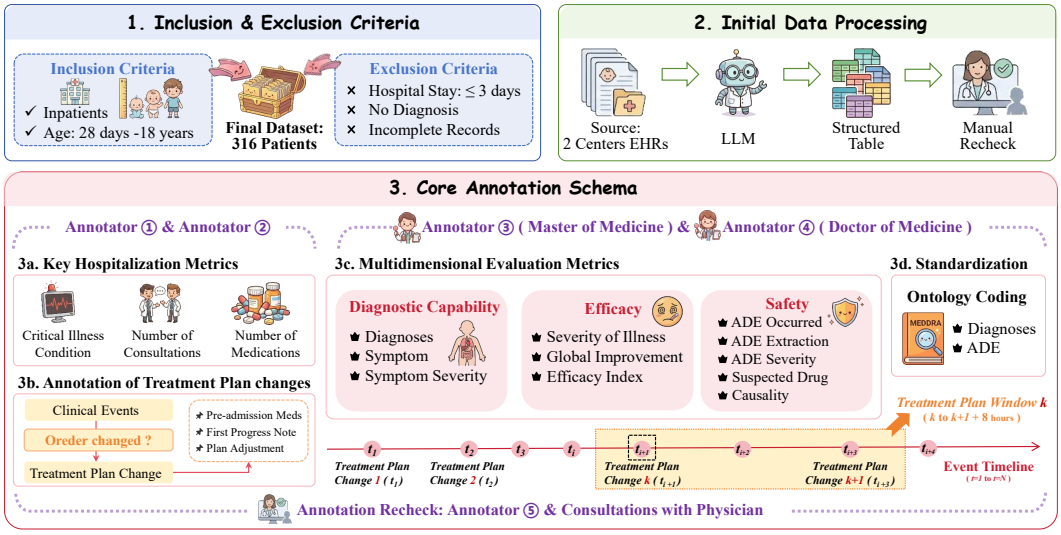


Figure 2: PedMed-Bench dataset construction and annotation pipeline. (1) Cohort selection with inclusion and exclusion criteria (§4.1); (2) LLM-assisted structuring and MedDRA normalization of de-identified EHRs from two centers, followed by manual verification (§4.1–4.2); (3) Core annotation of treatment trajectories spanning hospitalization metrics, multidimensional evaluation targets, and ontology coding (§4.3–4.5); (4) Double-blind quality assurance with third-party recheck (§4.6).

(GI), and Efficacy Index (EI); (3) **Medication Safety Prediction** predicts a tuple ($\mathcal{Y}_{safe}$) containing the suspected drug, adverse event name, Common Terminology Criteria for Adverse Events (CTCAE) grade, and causality grade.

Each regimen change is treated as a clinically meaningful decision point, and the subsequent Effective Observation Window captures delayed therapeutic and adverse outcomes. This window-based setup prevents hindsight bias from future records while preserving the core reasoning path from diagnosis to efficacy and medication-safety attribution (Harutyunyan et al., 2019; Singhal et al., 2023).

4 Pediatric Medication Benchmark

4.1 Data Acquisition and Cohort Selection

Data Collection. We curated a dataset of de-identified electronic health records (EHRs) from pediatric inpatients at two tertiary hospitals in China. From each center, we randomly sampled 500 candidate patients admitted between 2023 and 2024, then applied the inclusion and exclusion criteria below. EHRs comprise patient demographics, physician orders, clinical progress notes, and other relevant documents.

Inclusion Criteria. To build a representative benchmark, patients were included if they were hospitalized and aged between 28 days and 18 years, consistent with the International Council for Harmonisation E11 (ICH E11) guideline covering developmental stages from infancy to adolescence Food et al. (2000). Neonates younger than 28 days were excluded because their distinct physiological characteristics and drug metabolism substantially differ from those of older children.

Exclusion Criteria. To preserve benchmark validity and data quality, cases with a hospital stay of three days or less were removed, as short hospitalizations rarely provide sufficient longitudinal data to capture the dynamic relationship between medication adjustments and clinical progression. Patients discharged without a definitive diagnosis were excluded to avoid ambiguity in ground-truth labeling. Similarly, medical records lacking essential documentation, such as progress notes or physician orders, were discarded to maintain dataset integrity.

After this rigorous selection process, the final PedMed-Bench cohort consisted of 216 well-documented cases from Center 1 and 100 cases from Center 2, totaling 316 patients.

4.2 Data Standardization & Normalization

Data Structuring. The raw dataset comprised unstructured chronological text lacking standardized formatting, limiting its use for controlled multi-turn evaluation. To solve this, we developed an automated GPT-4o–based pipeline that converted progress notes and physician narratives into structured data. Following the Problem-Oriented Medical Record (POMR) framework Weed et al. (1968) and international documentation standards Soriano (2025), the dataset was organized into five dimensions: (1) History of Present Illness, (2) Clinical Trajectory, (3) Physical Examination, (4) Diagnostic Tests and Procedures, and (5) Physician Orders.

Medical Terminology Standardization and Coding. Variations in diagnosis and ADE descriptions across raw records and annotations stemmed from individual documentation habits, creating inconsistent synonyms and abbreviations. To ensure semantic consistency, we standardized all terms using Medical Dictionary for Regulatory Activities (MedDRA) v26.0 MedDRA et al. (2018), the international reference for regulatory communication and safety reporting. Each diagnosis and ADE was mapped to its lowest level term (LLT) and preferred term (PT), yielding unique, globally recognized identifiers for automated evaluation.

Human Recheck. To ensure accuracy, consistency, and completeness in the standardized data, a human-in-the-loop verification phase was implemented. GPT-4o was used only as an initial formatting and structuring assistant. All structured outputs were line-by-line verified against the original EHRs by domain experts. The final benchmark labels were determined entirely through expert annotation and adjudication. A pilot audit of GPT-4o-assisted structuring is provided in Appendix A.2.

Translation to English. To facilitate broader international participation and evaluation, all structured clinical records were professionally translated from Chinese to English. Translation followed medical documentation standards and was verified by bilingual clinical experts to preserve clinical semantics and terminology accuracy. Language-bias validation (Appendix D) confirms that the translated benchmark maintains equivalent difficulty and does not introduce systematic evaluation artifacts.

4.3 Dynamic Treatment Trajectory Extraction

Event Filtering and Dual-Sequence Construction. Clinical narratives encompass diverse inpatient activities such as ward rounds, examinations, consultations, and medical orders. To capture the decision dynamics underlying treatment modulation, we extracted pivotal nodes reflecting regimen changes from the complete event stream. This process produced a Dual-Stream Temporal Structure: (1) a comprehensive Inpatient Event Sequence tracing disease evolution, and (2) a Treatment Adjustment Sequence sparsely encoding physicians' iterative intervention strategies.

Taxonomy of Therapeutic Actions. Each identified regimen change was categorized into three decision phases spanning the therapeutic lifecycle: (1) Pre-admission Medication History (drugs administered prior to hospitalization), (2) Initial Inpatient Regimen (baseline treatment established at admission), and (3) Dynamic Regimen Adjustment (subsequent additions, discontinuations, or dosage modifications driven by disease progression).

Quality Control. Annotation employed a double-blind with adjudication protocol: two independent annotators identified adjustment events, followed by third-party review to ensure accuracy and completeness of final decision nodes.

4.4 Annotation of Medication Efficacy and Safety

Effective Observation Window. Before evaluating, we defined an Effective Observation Window spanning from treatment initiation (T_{start}) to 8 hours after the next regimen change (T_{next}). This window aligns with pediatric treatment conventions and pharmacokinetic (PK) principles Huang et al. (2012); Nelson et al. (2023), allowing detection of terminal therapeutic effects and delayed adverse events.

Task Annotation. For each identified regimen change, we conducted fine-grained annotations across three dimensions, symptomatology, efficacy, and safety, within the defined window. Comprehensive definitions and standards are provided in Appendix A.5.

- **Task 1: Symptomatology Diagnosis Effectiveness.** Progress notes and physical examination data at window onset yielded two metrics: (1) Symptoms, derived from patient-reported impressions; and (2) Symptom Severity, rated on a three-level scale.
- **Task 2: Therapeutic Efficacy Prediction.** Adapting the Clinical Global Impression (CGI) Guy (1976) scale for pediatric use, we assessed: (1) Severity of Illness (SI), describing disease status under the current regimen; (2) Global Improvement (GI), capturing relative improvement across the window; and (3) Efficacy Index (EI), integrating therapeutic benefit and adverse risk.
- **Task 3: Medication Safety Prediction.** Drug-related risks were assessed via 5 criteria: (1) ADE occurrence; (2) ADE description; (3) ADE severity, graded by CTCAE v6.0 and Hy's Law of Health et al. (2010); FDA (2009); (4) Suspected drug, identified through case history and literature evidence; and (5) Drug–ADE causality, evaluated using World Health Organization–Uppsala Monitoring Centre (WHO-UMC) Organization et al. (2005) and Roussel Uclaf Causality Assessment Method (RUCAM) Danan & Teschke (2015).

Quality Assurance: Annotation reliability was ensured through a double-blind hierarchical adjudication process: two clinical experts annotated independently, a third resolved disagreements, and complex cases were escalated to a multidisciplinary team of senior clinical pharmacists for final adjudication.

4.5 Clinical Complexity Annotation

To capture the multidimensional difficulty of cases in PedMed-Bench, we define three quantitative metrics: (1) Collaboration Complexity, based on the frequency of interdepartmental consultations, reflecting systemic involvement; (2) Treatment Complexity, measured by the total number of administered medications, indicating regimen intricacy; and (3) Critical Illness Status, determined by the issuance of a Critical Illness Notice as a benchmark for severity. We ensured consistent labeling through a double-blind annotation protocol: two clinically trained reviewers independently labeled all cases, with a third expert resolving disagreements.

4.6 Overall Quality Assurance

To ensure the reliability of the dataset, we implemented strict annotator qualification protocols and standardized procedures. Details are provided in Appendix A.7.

All candidate annotators completed a qualification test of 100 pilot samples. Their annotations were reviewed by domain experts, and only those achieving at least 80% accuracy were admitted to the formal annotation phase. The 80% threshold is consistent with established substantial-agreement criteria in medical annotation (Landis & Koch, 1977), dataset-curation practices that seek to control label noise (Liao et al., 2021), and reported expert agreement ranges in clinically ambiguous diagnostic tasks (Gulshan et al., 2016; Rajpurkar et al., 2018). The overall Cohen's Kappa among annotators was 0.74, indicating substantial agreement Landis & Koch (1977). Specifically, average Kappa values were 0.66 for the symptomatology dimension and 0.79 for medication safety. For medication efficacy, evaluated on an ordinal scale, we used Spearman's rank correlation, yielding an average of 0.78.

5 Evaluation & Metrics

This section systematically evaluates whether LLMs can accurately predict Pediatric Diagnosis Effectiveness, Therapeutic Efficacy, and Medication Safety. We report task-specific primary metrics for each evaluation dimension; detailed metric protocols, reference standards, and scoring definitions are provided in Appendix B.3 and Table 15.

Reporting conventions. Results are reported in percentage points. For entity-level tasks (diagnosis, symptom extraction, ADE extraction, suspected drug), we use strict matching against MedDRA-normalized gold labels and report F1 scores. For ordinal scales (symptom severity, SI, GI, EI, ADE severity, causality), we report exact agreement accuracy. SI, GI, and EI use Quadratic Weighted Kappa (QWK) normalized to $[-1, 1]$ then converted to percentage points; negative values indicate below-chance performance. "–" indicates cases where the model predicted no true positives for that metric. The best and second-best scores are highlighted in **green** and blue, respectively.

Table 2: Performance comparison on PedMed-Bench. See Section 5 for reporting conventions and metric definitions. Human performance is reported in Appendix B.4.

Model	Diagnostic Capability			Efficacy			Safety				
	Diag.	Symp. Ext.	Symp. Sev.	SI	GI	EI	ADE Occ.	ADE Ext.	ADE Sev.	Susp. Drug	Causality
Gemini 3.8 Flash	28.64	**51.26**	**67.95**	**54.08**	12.20	15.84	**74.63**	**12.89**	–	**12.90**	–
Gemini 3.1 Pro Preview	**31.92**	42.03	56.13	44.34	13.82	**34.17**	55.94	6.39	–	4.71	33.33
Gemini 3 Pro Preview	29.53	41.91	53.49	36.63	14.84	26.92	50.67	4.31	–	2.75	22.19
Claude Opus 5	15.64	17.67	43.99	35.88	0.18	14.99	31.65	2.69	–	2.34	**100.00**
GPT-5.4	22.23	26.13	58.79	35.69	13.57	24.51	23.25	2.16	–	1.05	53.92
DeepSeek V3.2	26.56	36.37	49.32	31.40	**30.81**	33.31	29.51	4.10	–	2.19	14.21
Qwen3-235B-A22B-Instruct	21.40	29.85	51.85	30.38	13.57	5.38	36.16	1.01	–	0.54	33.33
GPT-5	19.19	9.26	54.44	30.66	4.50	16.28	22.92	1.61	–	0.56	37.63
GPT-5.2	20.27	7.48	55.60	36.30	−0.65	22.24	31.39	2.10	–	1.19	0.90
Claude Sonnet 4.6	16.95	31.70	49.48	40.23	−19.08	22.53	44.35	2.44	**-10.00**	5.47	0.00
Qwen3-30B-A3B-Instruct	19.40	28.35	44.64	19.65	9.49	−3.24	27.28	1.66	–	1.37	0.00
Baichuan M3	24.32	46.07	56.04	34.12	8.83	−0.94	31.79	2.72	–	0.89	12.90
AntAngelMed	24.83	46.12	62.56	19.41	−8.25	16.99	47.34	3.30	–	1.49	0.00
MedGemma-4B-it	12.81	24.36	40.17	13.83	−4.47	1.67	27.84	0.85	–	0.30	0.00

Pediatric Diagnosis Effectiveness examines model competence in core pediatric diagnostic tasks: (1) ***Disease Diagnosis F1 (Diag.)***, measures the model's ability to identify the most plausible disease from a patient's clinical context (e.g., progress notes, examination findings), emphasizing integrated evidence reasoning. (2) ***Symptom Extraction F1 (Symp. Ext.)***, evaluates how reliably the model extracts patient-reported symptoms and clinician-observed signs from clinical narratives, reflecting clinical language understanding. (3) ***Symptom Severity Accuracy (Symp. Sev.)***, assesses the model's capacity to grade symptom intensity (e.g., mild, moderate, severe), supporting fine-grained patient state representation.

Pediatric Therapeutic Efficacy examines model accuracy in estimating treatment effectiveness through: (1) ***Severity of Illness Accuracy (SI)***, captures overall disease burden under treatment, reflecting global severity beyond individual symptoms. (2) ***Global Improvement Accuracy (GI)***, judges whether the model correctly characterizes patient trajectory over time (improvement, stability, or worsening). (3) ***Efficacy Index Accuracy (EI)***, integrates therapeutic benefits and treatment-related harms to quantify net efficacy, aligning evaluation with clinical benefit–risk reasoning.

Pediatric Medication Safety evaluates model performance in predicting medication-related risks across five dimensions: (1) ***ADE Occurrence Accuracy (ADE Occ.)*** measures the model's ability to detect any ADE during observation, serving as a global safety indicator. (2) ***ADE Extraction F1 (ADE Ext.)*** assesses precision in identifying specific adverse reaction types, supporting structured safety profiling and pharmacovigilance. (3) ***ADE Severity Accuracy (ADE Sev.)*** quantifies how accurately the model grades ADE seriousness, ensuring alignment with clinically relevant risk levels. (4) ***Suspected Drug F1 (Susp. Drug)*** evaluates whether the model correctly attributes causative medication(s) by integrating treatment history and contextual data. (5) ***Causality Judgement Accuracy (Causality)*** measures the model's ability to estimate causal strength between drugs and ADEs following pharmacovigilance principles.

6 Data Analysis

This study enrolled 316 pediatric inpatients across two tertiary hospitals, 216 at Center 1 and 100 at Center 2, to construct a benchmark dataset comprising 2,515 treatment nodes and 4,793 medical events. Statistical analysis shows that PedMed-Bench captures substantial demographic diversity, high clinical complexity, and distinctive temporal evolution. Due to space constraints, the detailed distribution proportions are provided in Figure 4 of the Appendix.

Real-world Patient Diversity. PedMed-Bench covers patients aged 28 days to 18 years, enabling evaluation of model generalization across developmental stages. The length of stay differs by center, with a median of 8 days at Center 1 and 5 days at Center 2, and 61.6% of patients male at Center 1 versus 53.0% at Center 2. These extended longitudinal records provide sufficient context for evaluating multi-turn clinical reasoning.

Therapeutic Complexity. The median number of concurrent medications per course is 10 at both centers. Additionally, 40.3% of cases at Center 1 and 33.0% of cases at Center 2 involved

interdepartmental consultations, demanding robust detection of drug–drug interactions and cross-disciplinary reasoning.

Acuity & Breadth. Critical illness accounts for 11.1% of cases at Center 1 and 8.0% at Center 2, providing cases for evaluating model behavior under high-acuity conditions in clinical decision-making. The dataset spans 186 unique clinical diagnoses and 69 unique adverse drug events at the MedDRA LLT level, covering multi-organ pathologies. Detailed distributions of diagnoses and adverse events are provided in Tables 16 and 17.

Dynamic Interaction Distribution. PedMed-Bench emphasizes temporal evolution in treatment decisions. The dataset comprises 2,515 treatment nodes across both centers, forming a dense, longitudinal decision sequence. As shown in Figure 5, diagnoses evolve markedly over time: admissions typically start with one to five diagnoses, while discharges frequently exhibit multimorbidity, reaching up to 8 diagnoses at Center 1 and up to 13 at Center 2. This progression mirrors the iterative refinement characteristic of clinical reasoning.

7 Experiments

7.1 Experimental Setup

We evaluate 14 representative LLMs: proprietary models (Claude, DeepSeek, Gemini, GPT series), open-source models (Qwen, Baichuan), and medical-domain models (MedGemma). The main experiment uses the full two-center cohort (316 patients, 858 treatment plans); medical-domain models and ablations use a single-center subset (121 patients, 330 plans). Details in Appendix B.

7.2 Main Results

PedMed-Bench remains challenging. As shown in Table 2, current LLMs achieve limited performance on several core pediatric medication evaluation tasks. The best model scores are 31.92 on diagnosis, 51.26 on symptom extraction, and 67.95 on symptom severity, indicating that longitudinal pediatric inpatient reasoning remains far from saturated.

Model strengths are task-specific. No model consistently dominates across all dimensions. Different models achieve the best scores on diagnosis, symptom extraction, symptom severity, SI, GI, EI, ADE occurrence, ADE extraction, suspected-drug identification, and causality. This suggests that PedMed-Bench evaluates multiple clinical abilities rather than a single medical knowledge score.

Medication safety is the key bottleneck. Models perform well on coarse ADE occurrence detection (best: 74.63), but fine-grained attribution remains challenging. ADE extraction and suspected-drug identification reach only 12.89 and 12.90, respectively. ADE severity classification performs below chance (-10.00), revealing systematic grading errors. Causality achieves 100.00 on a constrained subset, but the limited denominator highlights difficulty in joint ADE detection and causal reasoning.

7.3 Analysis

Temporal perturbations reveal shallow chronological grounding. We conduct time-series perturbations: temporal stripping, intra-group shuffling, and global shuffling. If models used chronological reasoning, corrupting temporal order should reduce performance. However, Table 3 shows ADE scores do not decrease, suggesting reliance on local drug–symptom co-occurrence rather than temporal ordering.

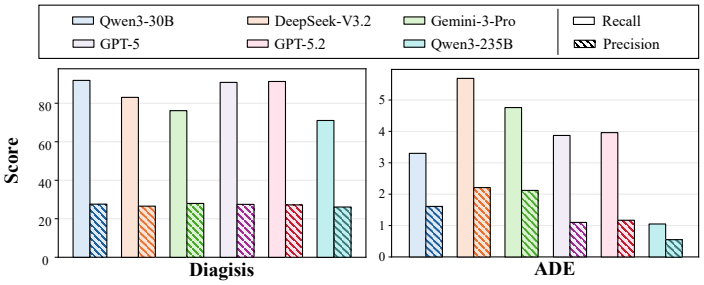
Counterfactual interventions reveal causal inertia. We remove suspected drugs or replace them with placebo interventions. Causally consistent models should reduce ADE predictions when medication evidence is removed. However, Table 3 shows ADE predictions remain high and even increase for DeepSeek-V3.2, suggesting weak sensitivity to medication-level causal evidence.

Safety failures arise from hierarchical bottlenecks. Oracle decomposition (Table 6) shows ADE occurrence alone is insufficient for attribution. Providing ADE names improves drug–event pairing, but even with both ADE and drug evidence, causality accuracy remains below 4%.

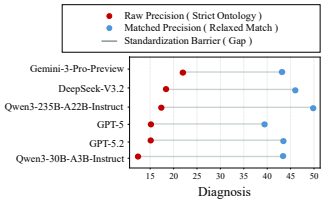
Diagnostic errors reflect over-prediction and label mismatch. Figure 3a shows models achieve higher recall than precision, indicating over-prediction. Figure 3b reveals a gap between strict MedDRA matching and semantic equivalence, suggesting some predictions are medically valid but not label-aligned.

Table 3: Perturbation analyses for ADE occurrence prediction. Time-series perturbation is evaluated on all treatment nodes ($N = 858$), while counterfactual medication intervention is evaluated on ADE-positive nodes with identifiable suspected drugs ($N = 206$). In the counterfactual setting, ADE Occ. denotes ADE-positive persistence after medication evidence is removed or substituted.

Model	**Time-series Perturbation** ($N = 858$)				**Counterfactual Intervention** ($N = 206$)		
	Baseline	**Strip**	**Intra**	**Global**	**Baseline**	**Full Removal**	**Placebo**
GPT-5.4	28.38	35.80	32.93	33.64	98.25	87.49	81.83
DeepSeek-V3.2	30.67	34.87	34.70	36.85	88.16	93.42	92.77



(a) Over-prediction behavior in diagnosis and symptom extraction.



(b) Gap between strict standardized-label matching and semantic equivalence.

Figure 3: Diagnostic error analysis. (a) Models show an over-prediction tendency, favoring broader output coverage at the expense of precision. (b) Some model predictions are medically equivalent to the gold diagnosis but fail strict matching against MedDRA-normalized labels, revealing a standardized-label alignment gap.

Pediatric-specific factors reveal non-uniform failure patterns. Performance varies substantially across age, consultation complexity, and medication count. Diagnostic accuracy peaks in infancy but drops in preschool groups, while safety metrics decline sharply in adolescents. Models show greater instability with increased polypharmacy and consultation complexity (Appendix C).

7.4 Exploration

We examine whether scaling, reasoning, prompting, retrieval, memory, and clinical prompting mitigate PedMed-Bench failures.

Scaling shows non-monotonic effects. Table 7 shows larger Qwen3 models do not yield uniform gains: diagnosis peaks at 4B, symptom severity improves at 235B, yet fine-grained safety remains below 1%.

Reasoning and prompting provide selective gains. Table 8 shows reasoning-oriented models improve several metrics, with DeepSeek-R1 leading on diagnosis, symptom extraction, ADE extraction, suspected-drug identification, and causality. Yet gains are task-dependent: Qwen3-235B-Thinking excels on symptom severity, Qwen3-30B-Thinking on EI, while ICL and CoT show limited benefits.

Retrieval, memory, and clinical prompting remain insufficient. RAG shows source-dependent effects: Google-RAG improves ADE occurrence for GPT-5, while PubMed-RAG improves efficacy metrics for GPT-5.4. LTM and CRP help some coarse metrics but often degrade fine-grained tasks.

8 Conclusion

We introduce PedMed-Bench, the first benchmark for pediatric medication reasoning from real-world inpatient EHRs. Grounded in international clinical standards (ICH E11, MedDRA, CGI, CTCAE, RUCAM), it evaluates LLMs across 11 metrics spanning diagnostic capability, therapeutic efficacy, and medication safety. Evaluation of 14 LLMs reveals: (1) no model consistently dominates, indicating distinct clinical reasoning abilities; (2) temporal perturbation and counterfactual interventions expose reliance on local co-occurrence rather than robust causal reasoning; (3) scaling, prompting, retrieval, and memory yield task-dependent gains but fail at fine-grained safety attribution (best ADE extraction F1: 12.89%). We hope PedMed-Bench catalyzes safer clinical AI for pediatric healthcare.

AI USE STATEMENT

We used generative AI tools in three ways, consistent with our disclosure at submission. First, for aiding and polishing writing: portions of this manuscript's prose were drafted or refined with the assistance of a large language model, and all AI-assisted text was reviewed and edited by the authors, who take full responsibility for its accuracy and claims. Second, for retrieval and discovery: we used AI-assisted literature search to help identify related work during the writing of Section 2 (Related Work); all cited works were subsequently verified by the authors against their original sources. Third, for generating synthetic datasets: GPT-4o was used to assist in structuring unstructured electronic health record narratives from both centers into the standardized clinical dimensions described in Appendix A.2; this tool served strictly as a formatting assistant and did not generate benchmark labels, which were determined entirely through expert annotation and adjudication (Appendix A.2). In a separate, auxiliary role, GPT-4o-mini was used as a semantic-equivalence judge in the supplementary diagnosis-matching analysis (Appendix C.6); this analysis does not contribute to the main benchmark score and was validated against blinded human evaluation. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involved a retrospective analysis of de-identified electronic health record data from pediatric inpatients at two tertiary hospitals in China. All patient- and physician-identifying information, including names, home addresses, national ID numbers, admission identifiers, and geographic references, was removed prior to data processing using a dual-mechanism approach combining regular expressions with manual review. As a retrospective analysis of de-identified data, this benchmark construction task is exempt from additional Institutional Review Board review under institutional guidelines. Annotators were recruited as senior pharmacists and clinically trained medical annotators; medical students participated only in preliminary extraction and terminology normalization after completing standardized training and qualification tests, while clinical pharmacists and senior experts adjudicated all efficacy and safety labels. All annotators demonstrated sufficient English reading proficiency for consulting international pharmacovigilance references, provided informed consent, and were compensated above the local minimum wage. To reduce geographical bias, annotators were blinded to the data source and regional information was masked. We publicly release only de-identified structured benchmark instances, prompt templates, and evaluation code; raw identifiable EHR text is not released. We emphasize that PedMed-Bench is intended as an evaluation resource and a diagnostic aid requiring mandatory human expert validation, and must not be used to provide direct prescription advice to patients.

REPRODUCIBILITY STATEMENT

To support reproducibility, we publicly release the de-identified structured benchmark instances, all prompt templates, and the full evaluation code at the anonymized project link given in Section 1. The released data contain no raw identifiable EHR text or personally identifiable information. Appendix A.2 documents the GPT-4o-assisted structuring workflow, including the prompt design, the human-in-the-loop verification protocol, and a pilot audit reporting that approximately 97% of structuring differences were formatting or linguistic and about 3% were factual discrepancies, all corrected before release. Sections A.3 through A.5 specify the annotation standards, reference scales, and scoring definitions for each task dimension. Appendix B provides the complete experimental protocol, including model identities and parameter sizes (Table **??**), access dates, the uniform treatment-node prediction setting, input construction from information available strictly before the prediction point, prompt and parsing procedures, and the formal definitions of all task-specific metrics. We additionally report inter-annotator reliability statistics in Table 4, and provide a human-reference baseline alongside all model results to contextualize performance. For the auxiliary semantic-matching analysis, we document the judge model, its validation against blinded human evaluation on 363 sampled pairs, and a cross-model robustness check with two alternative judges in Appendix C.6.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Anthropic. *Introducing Claude Sonnet 4.6*, February 2026a. URL https://www.anthropic.com/news/claude-sonnet-4-6. Anthropic News.

Anthropic. Introducing claude opus 5. https://www.anthropic.com/news/claude-opus-5, 2026b. Accessed: 2026-09-26.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 2025.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Sharon Conroy, Imti Choonara, Piero Impicciatore, Angelika Mohn, Henrik Arnell, Anders Rane, Carmen Knoeppel, Hannsjoerg Seyberth, Chiara Pandolfini, Maria Pia Raffaelli, et al. Survey of unlicensed and off label drug use in paediatric wards in european countries. *bmj*, 320(7227):79–82, 2000.

Gaby Danan and Rolf Teschke. Rucam in drug and herb induced liver injury: the update. *International journal of molecular sciences*, 17(1):14, 2015.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, et al. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*, 2026.

US FDA. Guidance for industry drug-induced liver injury: premarketing clinical evaluation. *Silver Spring: US FDA*, 2009.

Chris Feudtner, Dingwei Dai, Kari R Hexem, Xianqun Luan, and Talene A Metjian. Prevalence of polypharmacy exposure among hospitalized children in the united states. *Archives of pediatrics & adolescent medicine*, 166(1):9–16, 2012.

US Food, Drug Administration, et al. E11 clinical investigation of medicinal products in the pediatric population. *Access at: https://www. fda. gov/regulatory-information/search-fda-guidance-documents/e11-clinical-investigation-medicinal-products-pediatric-population (accessed on April 12, 2020)*, 12362934, 2000.

Google DeepMind. *Gemini 3.1 Pro*, May 2026a. URL https://deepmind.google/models/gemini/pro/. Google DeepMind Models.

Google DeepMind. Gemini 3.8 flash. https://deepmind.google/models/gemini/flash/, 2026b. Accessed: 2026-09-26.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

William Guy. *ECDEU assessment manual for psychopharmacology*. US Department of Health, Education, and Welfare, Public Health Service . . . , 1976.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Lorna Hazell and Saad AW Shakir. Under-reporting of adverse drug reactions. *Drug safety*, 29(5): 385–396, 2006.

Shiew-Mei Huang, Juan JL Lertora, and Arthur J Atkinson Jr. *Principles of clinical pharmacology*. Academic Press, 2012.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9): AIdbp2500144, 2025.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86, 2025.

Rainu Kaushal, David W Bates, Christopher Landrigan, Kathryn J McKenna, Margaret D Clapp, Frank Federico, and Donald A Goldmann. Medication errors and adverse drug events in pediatric inpatients. *Jama*, 285(16):2114–2120, 2001.

Gregory L Kearns, Susan M Abdel-Rahman, Sarah W Alander, Douglas L Blowey, J Steven Leeder, and Ralph E Kauffman. Developmental pharmacology—drug disposition, action, and therapy in infants and children. *New England Journal of Medicine*, 349(12):1157–1167, 2003.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, et al. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745, 2024.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. Towards good practices for efficiently annotating large-scale image classification datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4350–4359, 2021.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, Peter Brodeur, et al. Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120, 2025.

MSSO MedDRA et al. Introductory guide for standardised meddra queries (smqs) version 21.0. *Geneva, Switzerland: International Federation of Pharmaceutical Manufacturers and Associations*, 2018.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*, 2025.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

John D Nelson, American Academy of Pediatrics, et al. *2023 Nelson's Pediatric Antimicrobial Therapy*. American Academy of Pediatrics, 2023.

US Department of Health, Human Services, et al. National cancer institute. *Common terminology criteria for adverse events (CTCAE) version*, 4:1–194, 2010.

OpenAI. *GPT-4o-mini: Advancing Cost-Efficient Intelligence*, July 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient\-intelligence/. OpenAI Product Announcement.

OpenAI. *Introducing GPT-5*, September 2025a. URL https://openai.com/index/introducing-gpt-5/. OpenAI Research Index.

OpenAI. *Introducing GPT-5.2*, December 2025b. URL https://openai.com/index/introducing-gpt-5-2/. OpenAI Product Launch Announcement.

OpenAI. Introducing gpt-5.4, March 5 2026. URL https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/. Accessed: September 26, 2026.

World Health Organization et al. The use of the who-umc system for standardized case causality assessment. *Uppsala: The Uppsala Monitoring Centre*, 1, 2005.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Rainier P Soriano. *Bates' Guide To Physical Examination and History Taking*. Lippincott Williams & Wilkins, 2025.

AntAngelMed Team. Antangelmed: A high-performance medical language model with efficient moe-powered clinical reasoning, 2025. URL https://huggingface.co/MedAIBase/AntAngelMed.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Lawrence L Weed et al. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600, 1968.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Xin Sun, Ya Zhang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *arXiv preprint arXiv:2506.20430*, 2025.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, 2025.

Siyu Zhu, Yue Xie, Yongyu Tang, Zhikang Yu, Ran Zhao, and Xiaoyan Dong. New chapter in pediatric medicine: technological evolution, application, and evaluation system of large language models. *European journal of pediatrics*, 184(12):809, 2025a.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*, 2025b.

Hui Zong, Jiaxue Cha, Jiao Wang, Yu Song, Yan Zhao, Muyun Shi, and Bairong Shen. A dataset for evaluating large language models on chinese national medical licensing examinations. *Scientific Data*, 2026.

APPENDIX

A DATASET ANNOTATION DETAILS

A.1 THE DETAILS OF DATA ACQUISITION AND COHORT SELECTION

Data Collection We extracted raw Electronic Health Records (EHRs) data from the pediatric departments of two tertiary hospitals in China, covering the period from 2023 to 2024. The initial candidate pool contained the complete hospitalization trajectories of 500 pediatric patients per center (1,000 in total). To construct PedMed-Bench, we performed a deep integration of the following three key categories of clinical data:

- **Structured Demographics:** Including age, gender, weight, and vital signs upon admission.
- **Semi-structured Progress Notes:** Encompassing daily ward round records, medication adjustments, disease evolution notes, and multi-disciplinary team (MDT) consultation advice.
- **Unstructured Clinical Narratives:** Extracted from physician notes, including oral records such as symptom descriptions, differential diagnoses, and clinical analyses.

Privacy Protection and De-identification Prior to the data cleaning phase, all raw records were anonymized in strict accordance with hospital data security protocols. We developed a dual-mechanism approach combining regular expressions with manual review to remove all patient and physician names, home addresses, national ID numbers, and specific hospital admission IDs. Furthermore, we masked all terms related to geographic locations to ensure clinical information was preserved while eliminating regional privacy risks.

Rigorous Enrollment and Screening Logic To ensure that PedMed-Bench effectively challenges LLMs in complex pediatric decision-making, we implemented a multi-stage screening workflow. Regarding inclusion criteria, we prioritized full clinical course coverage by requiring all subjects to be inpatients; compared to outpatient data, inpatient records offer a significantly longer observation window to fully capture the medication adjustment process. Furthermore, to ensure broad physiological development coverage, the age range was restricted to 28 days to 18 years. Neonates (< 28 days) were excluded based on clinical consensus regarding their unique blood-brain barrier development and hepatic/renal metabolism, thereby aligning our scope with the most general pediatric pharmacotherapy guidelines as defined by ICH E11. Conversely, our exclusion criteria focused on ensuring data integrity and evaluative validity. We first excluded cases with a length of stay ≤ 3 days due to insufficient temporal depth, as records under 72 hours typically fail to capture the dynamic evolution of drug efficacy or the identification logic for ADEs. Additionally, cases lacking a confirmed diagnosis upon discharge were removed to ensure a definitive Ground Truth for objective quantitative evaluation. Finally, we filtered out fragmented cases with mismatched progress notes and physician orders, preventing models from generating "hallucinated" diagnoses due to incomplete clinical information.

Final Dataset Composition Following the stratified screening process, we finalized 316 core cases (216 from Center 1, 100 from Center 2). This ensures that PedMed-Bench is not merely a simple Q&A set, but a dynamic evaluation benchmark characterized by high information density and complex medical reasoning paths.

A.2 THE DETAILS OF DATA STRUCTURAL NORMALIZATION

Automated Structuring Workflow with GPT-4o Given that the raw Electronic Health Records (EHRs) were stored as unstructured and fragmented temporal sequences, we developed an automated pipeline leveraging `GPT-4o` to map these clinical narratives into a standardized format. The process reconfigures discrete progress notes and physician entries into five core clinical dimensions: (1) History of Present Illness (HPI), (2) Clinical Course Status, (3) Physical Examination (PE), (4) Laboratory and Imaging Tests, and (5) Physician Orders. This hierarchical architecture was strictly established based on the Problem-Oriented Medical Record (POMR) framework proposed by Weed (1968) and aligns with international standards for inpatient documentation.

To minimize hallucinations and maximize the preservation of medical semantics, we employed the following prompting strategy:

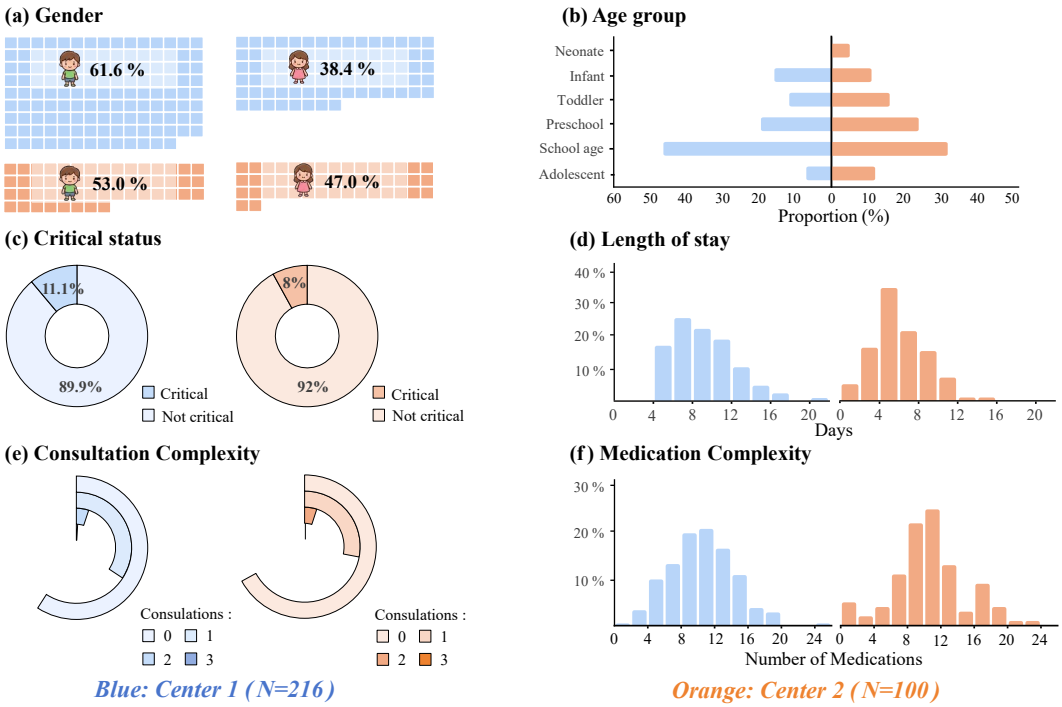


Figure 4: **PedMed-Bench dataset statistics across two centers.** Blue bars represent Center 1 ($N = 216$) and orange bars represent Center 2 ($N = 100$). (a) **Gender**: Center 1 has 61.6% male vs. 38.4% female; Center 2 has 53.0% male vs. 47.0% female. (b) **Age group**: patient distribution across six developmental stages from neonate to adolescent. (c) **Critical status**: 11.1% critical cases in Center 1 and 8% in Center 2. (d) **Length of stay**: distribution of hospital stay duration in days. (e) **Consultation complexity**: distribution of multidisciplinary consultation counts per patient. (f) **Medication complexity**: distribution of concurrent medication counts per course, reflecting polypharmacy burden.

Table 4: Inter-rater Reliability (Kappa and Spearman Coefficients) for Clinical Annotations.

Dimension	Metric	Reliability Metric	Value	Avg. Reliability
Diagnostic Capability	Symp. Ext.	Cohen's Kappa	0.7410	0.6901
	Symp. Sev.	Cohen's Kappa	0.6392	
Efficacy	SI	Spearman Correlation	0.7506	0.8293
	GI	Spearman Correlation	0.8038	
	EI (Eff.)	Spearman Correlation	0.7903	
	EI (Side)	Spearman Correlation	0.9725	
Safety	ADE Occ.	Cohen's Kappa	0.9632	0.8081
	ADE Ext.	Cohen's Kappa	0.9209	
	ADE Sev.	Cohen's Kappa	0.8654	
	Susp. Drug	Cohen's Kappa	0.6729	
	Causality	Cohen's Kappa	0.6181	
Overall	—	—	—	**0.7758**

[System Prompt]
You are a medical information extraction expert with extensive experience in pediatrics. Your task is to convert unstructured pediatric progress notes into a standardized JSON format based on the Weed (1968) POMR framework.

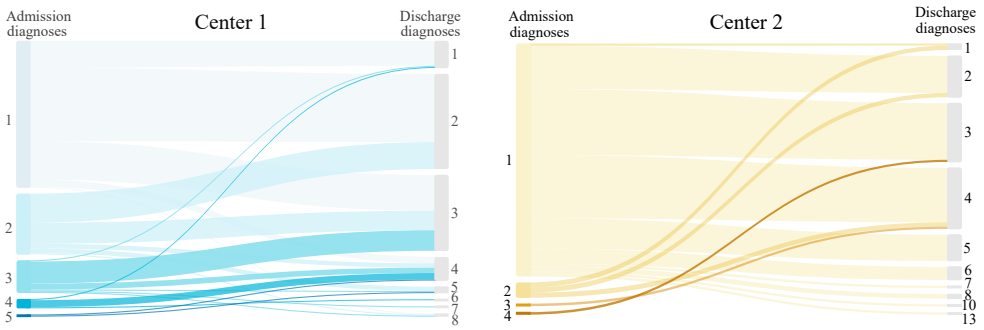


Figure 5: **Diagnosis evolution from admission to discharge across two centers.** Each Sankey diagram traces how admission diagnoses flow into discharge diagnoses for Center 1 ($N = 216$, left) and Center 2 ($N = 100$, right). The widths of the flows are proportional to patient counts. In both centers, patients typically present with few diagnoses at admission, reaching up to 5 in Center 1 and up to 4 in Center 2, but frequently discharge with a greater number and variety of diagnoses, with discharge diagnosis counts reaching up to 8 in Center 1 and up to 13 in Center 2. This progressive multimorbidity reflects the dynamic nature of pediatric inpatient care, where disease states evolve in response to successive medication adjustments, highlighting the need for benchmarks that capture longitudinal treatment-outcome reasoning rather than static single-step prediction.

[Instructions]

- **Atomic Extraction**: Decompose the text into minimal clinical semantic units to ensure no information overlap between dimensions.
- **Temporal Alignment**: Maintain the timestamps of the original records. If a time is ambiguous, mark it as "Unknown."
- **Fidelity Principle**: Subjective inference is strictly prohibited. Extract only explicitly documented data (e.g., "Body temperature 38.5°C" rather than "fever").

[Output Schema]

```
{ "HPI": [...], "Status": [...], "Physical_Ex": [...],
"Lab_Tests": [...], "Orders": [...] }
```

Human-in-the-Loop Verification and Expert Instructions To ensure the integrity and accuracy of the generated data, we implemented a rigorous Human-in-the-Loop (HITL) verification mechanism. Medical experts conducted a line-by-line audit comparing the structured output against the original EHRs for all 316 core cases. The expert review instructions focused on three core objectives: **completeness, accuracy, and semantic consistency**. Specifically, experts were tasked with identifying any omissions of critical medical information, such as dosage adjustments for high-risk medications or life-threatening laboratory critical values. Furthermore, experts audited the structured descriptions to ensure that clinical intent remained undistorted—for instance, verifying that a "suspected viral infection" was not erroneously extracted as a "confirmed diagnosis." Any significant semantic deviations identified in the `GPT-4o` output were flagged as "Critical Errors" and manually corrected. Each manual correction was subsequently reviewed and confirmed by a third expert. This intensive manual oversight ensured that the final dataset achieved the highest precision, providing a high-fidelity foundation for the PedMed-Bench benchmark.

Audit of GPT-4o-assisted Structuring GPT-4o was used as a structuring assistant rather than as an annotator. To assess possible structuring bias, we compared GPT-4o-assisted outputs with the final expert-verified structured records. The pilot audit showed that approximately 97% of the differences were formatting or linguistic adjustments, while about 3% involved factual discrepancies such as missing temporal qualifiers, incomplete medication details, or over-normalized clinical expressions. All factual discrepancies were manually corrected and confirmed by experts before benchmark construction. Therefore, GPT-4o served only as an efficiency tool, while the final benchmark reference was produced through human expert verification. The 3% factual discrepancy rate was

Table 5: Pilot audit of GPT-4o-assisted EHR structuring. Time and cost estimates are scaled to the initial 500-patient-per-center candidate pool. All GPT-4o outputs were verified line-by-line by human experts before label construction.

Efficiency comparison			
Metric	**Manual Annotation**	**GPT-4o Assisted**	**Improvement**
Time per patient	∼22 min 22 sec	∼15 sec	∼89× faster
Total time for 1,000 candidate patients	∼75 hours	∼50 min	∼89× faster
Total cost	∼$620	∼$30	∼20× lower
Human audit of GPT-4o-assisted outputs			
Audit item	**Observed result**		**Action**
Formatting / linguistic differences	97%		Accepted after review
Factual discrepancies	3%		Corrected by experts
Final benchmark labels	Expert-verified reference		Human adjudication

measured before expert correction and therefore does not reflect residual errors in the released benchmark labels.

A.3 The Details of Clinical Complexity Quantification and Stratification

Quantitative Metrics To comprehensively evaluate the management difficulty within *PedMed-Bench*, we defined three core metrics for stratified quantification:

- **Collaboration Complexity**: Measured by the frequency of unique departmental consultations. This metric excludes consultation requests without official reports or those declined by guardians, reflecting the breadth of systemic involvement and the necessity for interdisciplinary synergy.
- **Treatment Complexity**: Calculated as the total number of distinct medications used throughout the clinical course. The scope includes pre-hospitalization medications administered by parents while strictly excluding medical devices (e.g., cooling patches) or non-pharmacological treatments (e.g., acupuncture).
- **Critical Illness Status**: Determined by the explicit issuance of a "Critical Illness Notice" in the medical record, serving as the definitive gold standard for high clinical acuity.

Quality Control We implemented a **Double-blind Annotation with Adjudication** mechanism to ensure data objectivity and fidelity:

- **Independent Labeling**: Two annotators with clinical backgrounds performed independent, back-to-back labeling according to the standardized instructions.
- **Expert Adjudication**: All discrepancies were resolved by a senior expert to maintain high inter-rater reliability.

Instruction:

[Task 1: Collaboration Complexity (Consultation Count)]

- **Inclusion Criteria**: Only consultations formally requested by the attending physician AND supported by a recorded consultation report in the progress notes are counted.
- **Exclusion Criteria**: Consultations requested but missing records, or those rejected by guardians, are excluded.
- **De-duplication**: Multiple consultations from the same department for the same patient count as ONE unique consultation.

[Task 2: Treatment Complexity (Polypharmacy)]

- **Scope**: Include all medications used throughout the entire course, including pre-hospitalization medications administered by parents.
- **Drug Identification**: All names must match the provided "Drug Package Insert List." If a drug's specific name is missing (e.g., recorded only as "antipyretic"), label it by its category (e.g., "Antipyretics").
- **Exclusions**: Medical devices, physical therapies, and non-pharmacological treatments (e.g., cooling patches, acupuncture, massage/Tuina) are NOT counted as medications.

[Task 3: Critical Illness Status (Acuity)]

- **Gold Standard**: A case is marked as "Critical" (1) if and only if a "Critical Illness Notice" (Critical Condition Notice) is explicitly documented in the medical record. Otherwise, mark as 0.

A.4 The Details of Dynamic Treatment Trajectory Extraction

Event Filtering and Dual-Sequence Construction Inpatient progress notes contain heterogeneous medical events, such as ward round records, laboratory results, consultation opinions, status updates, and medication revisions. To focus on the dynamic decision-making process of medication adjustments, we precisely extract key decision nodes involving "Treatment Plan Changes" from the redundant inpatient event stream. This filtering process establishes a dual temporal structure to describe the patient's journey:

- **Complete Inpatient Event Sequence**: A full chronological record of all medical events, used for a panoramic description of the disease evolution trajectory.
- **Treatment Plan Adjustment Sequence**: A sparse sequence consisting solely of key events involving medication changes, used to characterize the physician's dynamic intervention strategies.

Classification of Clinical Actions and Annotation Guidelines Extracted changes in treatment plans are strictly categorized into three decision types to cover the complete therapeutic cycle: (1) Pre-hospitalization Medication History, (2) Initial Treatment Plan, and (3) Treatment Plan Adjustments. To ensure high-quality extraction, we implemented the following specialized annotation instructions:

Instruction:

[Task: Treatment Plan Changes]

- **Primary Focus**: Revisions to pharmacological therapy, encompassing Drug Name, Frequency, or Dosage.
- **Inclusions**: Adjustments due to fluid therapy, drug substitutions (e.g., pharmacy out-of-stock), and the use of self-provided medications by parents.
- **Exclusions**: Changes in nursing levels, supplementary lab/imaging requests, or rehabilitation therapies (e.g., acupuncture, Tuina, cooling patches) do NOT constitute a treatment plan change.

Quality Control We continued to employ the **Double-blind Annotation with Adjudication** mechanism for this section. Two annotators independently identified and extracted medication change events according to the guidelines above, which were then reviewed by a third expert to ensure the completeness and accuracy of the extracted decision nodes.

A.5 The Details of Annotation of Medication Efficacy and Safety

To ensure the authority and high fidelity of the benchmark data, we implemented a rigorous "double-blind annotation with hierarchical adjudication" workflow. Details regarding the annotation of metrics are provided in Table 15. Notably, the EI (Efficacy Index) is calculated as the quotient of two metrics: the therapeutic efficacy assessment and the side effect evaluation.

Prior to the formal labeling phase, all annotators underwent systematic training based on official drug package inserts and clinical scale standards, and were required to pass a validation test consisting of 100 representative samples. During the implementation stage, two experts with clinical backgrounds performed independent, back-to-back labeling. General discrepancies were directly adjudicated by a third senior expert. For highly controversial and complex cases, an upgraded arbitration mechanism was triggered, involving a Multi-disciplinary Team (MDT) of senior clinical pharmacists to determine the final Ground Truth from a professional pharmacological perspective.

Instruction:

[1. Effective Observation Window]
The evaluation must cover the period from T_{start} to $T_{next} + 8h$. This buffer ensures the capture of residual drug effects.

[2. ADE Identification Logic]
An event is labeled as an ADE if: (1) It was not present upon admission; (2) It occurred within the observation window; (3) It is documented in the provided official drug labeling.

[3. Consensus and Escalation]
Double-blind labeling is mandatory. Discrepancies go to a third reviewer. Unresolved complex cases trigger an MDT discussion with clinical pharmacists.

A.6 The Details of Medical Terminology Standardization and Coding

In accordance with the MedDRA 26.0 standard, all extracted "clinical diagnoses" and "ADEs" were mapped to the Lowest Level Terms (LLTs) and Preferred Terms (PTs) levels. This mapping process ensures that each clinical entity is assigned a globally unique standard code, significantly facilitating subsequent automated model evaluation.

The terminology coding also strictly adhered to the "double-blind independent annotation + expert adjudication" workflow. Two annotators independently completed the terminology mapping using the provided comprehensive MedDRA 26.0 coding table, followed by a review by a third expert. This review primarily focused on resolving the normalization of ambiguous terms to ensure coding accuracy. Following MedDRA standardization, the statistical distributions of clinical diagnoses and ADEs in the PedMed-Bench dataset are summarized in the tables 16 and 17.

Instruction:

[Objective]
The primary goal of this task is to eliminate semantic ambiguity and heterogeneity in pediatric electronic health records (EHR) caused by varied clinical writing habits. Annotators are required to map all identified “Clinical Diagnoses” and “Adverse Drug Events (ADE)” to international standard levels: **LLT** (Lowest Level Term) and **PT** (Preferred Term) using the MedDRA v26.0 database.

[Core Coding Logic]
Annotators must strictly adhere to the “Specificity” and “Uniqueness” principles during the mapping process:

- **LLT Identification**: Search for the most granular term in the coding table that matches the verbatim clinical description.
- *Example*: If a record states “Allergic Rash,” select the specific LLT “Allergic Rash” rather than the broader category “Rash”.
- **PT Mapping**: Every selected LLT must be linked to its medically validated PT. Ensure both the unique PT code and name are recorded to facilitate subsequent data aggregation.

[Clinical Execution Steps]

Table 6: Oracle bottleneck decomposition for medication safety prediction. We provide different oracle signals and evaluate whether models can recover downstream safety outputs. ADE–Drug Pair F1 denotes the F1 score of correctly identifying the adverse event and its suspected drug as a pair. Metrics corresponding to oracle-provided fields are omitted to avoid tautological evaluation. All scores are reported in percentage points.

Model	Oracle Setting	ADE Ext.	ADE–Drug Pair F1	Causality
DeepSeek-V3.2	Full prediction	3.07	0.95	0.93
	Oracle ADE Occ.	0.84	0.21	0.00
	Oracle ADE Name	22.97	8.63	1.81
	Oracle Drug	0.59	0.44	0.47
	Oracle ADE + Drug	–	–	3.74
GPT-5.4	Full prediction	1.93	0.67	1.75
	Oracle ADE Occ.	0.64	0.26	0.12
	Oracle ADE Name	20.28	8.71	1.64
	Oracle Drug	0.89	0.74	0.23
	Oracle ADE + Drug	–	–	3.09

- **Extraction**: Precisely identify clinical entities from progress notes, discharge summaries, or specialist consultations.
- **Search Strategy**: Query the MedDRA v26.0 terminology. Prioritize full terms that include specific anatomical sites or pathological features.
- **Mapping Scenarios**:
  - *Exact Match*: The verbatim clinical text aligns directly with an existing dictionary term.
  - *Synonym Conversion*: Convert clinical shorthand (e.g., "HF") or non-standard abbreviations into standardized medical terms (e.g., "Heart failure") before coding.
  - *Compositional Coding*: Deconstruct descriptions containing multiple distinct entities (e.g., "Flu with fever") into separate LLT/PT entries.

[Uncertainty and Arbitration]
In cases where the textual information in the EHR is vague or insufficient:

- Evaluate supplementary information, such as laboratory results or imaging reports, to formulate a more precise clinical rationale for the term.
- If a match still cannot be determined with certainty, mark the entry as **"To be Arbitrated"**. Do not perform subjective or speculative mapping.

A.7 The Details of Quality Assurance

To ensure the high fidelity of PedMed-Bench, we implemented a rigorous annotator qualification test (consisting of 100 samples, requiring an accuracy rate $\geq 80\%$) and a "double-blind independent annotation + expert adjudication" workflow. Statistical analysis shows that the overall Cohen's Kappa coefficient among annotators reached 0.74, indicating substantial agreement. Specifically, the medication safety dimension demonstrated the most robust performance (Avg. $\kappa = 0.79$), with the identification of ADE occurrence achieving near-perfect agreement ($\kappa = 0.97$). Furthermore, the average Spearman correlation coefficient for efficacy metrics reached 0.78. These data support the reliability of the annotated reference labels; detailed metric results are provided in Table 4.

Table 7: Model scaling analysis on the Qwen3 family. Results are reported using task-specific primary metrics in percentage points (%). SI, GI, and EI are reported as QWK normalized to $[-1, 1]$ and then converted to percentage points; negative values indicate performance below chance. The best and second-best performances in each column are highlighted in **green** and blue, respectively. Ties are highlighted with the same color. "–" denotes that no true-positive hits were found for the model, making the metric undefined.

Model	Diag.	Symp. Ext.	Symp. Sev.	SI	GI	EI	ADE Occ.	ADE Ext.	ADE Sev.	Susp. Drug	Causality
Qwen3-0.6B	11.17	27.78	35.42	−7.77	11.61	2.70	**72.15**	0.00	–	0.00	–
Qwen3-1.7B	14.74	29.09	35.67	5.84	2.30	7.68	57.66	0.00	–	0.00	–
Qwen3-4B	**22.15**	**35.40**	46.51	9.37	−1.23	0.22	55.65	**4.43**	**100.00**	**4.32**	0.00
Qwen3-8B	20.59	33.79	47.24	8.46	5.25	−0.58	52.33	2.66	0.00	2.48	–
Qwen3-14B	21.56	31.75	46.63	10.78	−1.72	2.28	33.03	1.44	50.00	1.11	33.33
Qwen3-30B-a3b-instruct	18.91	28.08	47.22	3.79	8.12	0.04	29.34	0.35	75.00	0.06	0.00
Qwen3-32B	22.08	32.84	43.56	12.07	19.87	**13.07**	39.99	1.76	**100.00**	1.65	–
Qwen3-235B-a22b-instruct	20.72	29.79	**54.97**	**18.86**	**9.88**	3.90	36.34	1.07	45.45	0.36	**40.00**

Table 8: Method exploration on PedMed-Bench. Results are reported using task-specific primary metrics in percentage points. Google-RAG denotes web search-based retrieval, while PubMed-RAG denotes academic biomedical literature retrieval. The best and second-best performances in each column are highlighted in **green** and blue, respectively.

Method	Diag.	Symp. Ext.	Symp. Sev.	SI	GI	EI	ADE Occ.	ADE Ext.	ADE Sev.	Susp. Drug	Causality
Reasoning-oriented models											
DeepSeek-R1	**24.92**	**44.11**	51.76	17.37	24.13	21.45	30.08	**3.03**	4.08	**1.21**	**2.33**
DeepSeek-V3.2-Thinking	23.50	39.25	49.69	18.53	20.16	12.24	29.16	2.43	3.50	1.03	1.57
Qwen3-235B-Thinking	24.13	43.63	**64.83**	29.60	19.46	20.86	29.74	1.78	1.86	0.93	1.28
Qwen3-30B-Thinking	19.60	37.10	47.60	20.05	20.40	**30.77**	40.60	0.12	0.12	0.00	0.00
Prompting strategies on GPT-5											
GPT-5	18.67	18.05	27.70	7.28	19.05	17.43	29.44	1.69	3.29	0.34	0.65
GPT-5 + CoT	14.84	17.50	31.26	7.23	20.12	19.88	28.20	1.68	3.10	0.36	1.13
GPT-5 + ICL	8.58	39.37	27.43	**68.47**	21.36	16.20	28.04	2.56	**4.58**	0.42	0.71
GPT-5 + Google-RAG	5.27	17.65	24.46	5.65	12.25	19.38	46.18	0.97	2.24	0.21	0.44
Retrieval, memory, and clinical prompting on GPT-5.4											
GPT-5.4	19.28	24.75	61.13	19.93	24.13	21.45	28.38	1.93	3.15	0.67	1.75
GPT-5.4 + PubMed-RAG	15.12	23.25	43.88	9.09	**27.62**	24.83	28.98	1.50	3.03	0.57	1.63
GPT-5.4 + LTM	12.53	20.44	39.78	15.27	19.23	1.86	34.92	0.24	0.47	0.06	0.12
GPT-5.4 + Google-RAG	14.21	23.44	43.45	9.21	27.16	24.13	27.95	1.44	3.03	0.54	1.46
GPT-5.4 + CRP	12.80	25.12	41.54	21.33	12.47	1.40	**47.65**	0.19	0.58	0.03	0.00

B Experimental Details

B.1 Model Selection

To comprehensively evaluate the clinical reasoning capabilities of LLMs in pediatric medication scenarios, we select a diverse set of representative models, covering proprietary frontier models, open-source general models, Chinese-domain models, and medical-domain models. In addition, we include human reference performance to contextualize the difficulty of *PedMed-Bench*.

Proprietary Models. We evaluate representative models from OpenAI, Anthropic, and Google. From OpenAI, we include **GPT-5** OpenAI (2025a), **GPT-5.2** OpenAI (2025b), and **GPT-5.4** OpenAI (2026), which represent recent general-purpose reasoning-oriented models. From Anthropic, we evaluate **Claude-4.6-Sonnet** Anthropic (2026a) and **Claude-Opus-5** Anthropic (2026b), which are widely used for long-form instruction following and clinical text reasoning. From Google, we include **Gemini-3-Pro-Preview**, **Gemini-3.1-Pro-Preview** Google DeepMind (2026a), and **Gemini-3.8-Flash** Google DeepMind (2026b) to assess models with strong long-context capabilities and rapid response performance.

Open-source and Domain-specific Models. We further evaluate several open-source and domain-adapted models. **DeepSeek-V3.2** Liu et al. (2025) is included as a strong general-purpose open model with competitive reasoning performance. **Qwen3-235B-Instruct** Yang et al. (2025) and **Qwen3-30B-Instruct** Yang et al. (2025) are selected to represent large-scale and more efficient

open-weight systems, respectively. For medical-domain models, we evaluate the Chinese medical model **Baichuan-M3** Dou et al. (2026), **MedGemma-4B-it** Sellergren et al. (2025) as a lightweight medical-domain model, and **AntAngelMed** Team (2025) as a high-performance medical language model with efficient MoE-powered clinical reasoning, to assess whether medical specialization improves pediatric medication evaluation.

This model set allows us to compare general-purpose reasoning ability, open-source scalability, Chinese clinical language adaptation, and medical-domain specialization under the same treatment-node prediction protocol.

Model Parameters. Proprietary models (GPT, Claude, Gemini series) do not disclose parameter counts. Among open-weight models, DeepSeek-V3.2 uses 685B total parameters with 37B active per forward pass (MoE architecture), Qwen3-235B-A22B-Instruct uses 235B total with 22B active (MoE), and Qwen3-30B-A3B-Instruct uses 30B total with 3B active (MoE). Dense Qwen3 models range from 0.6B to 32B parameters. Medical-domain models include Baichuan-M3 (undisclosed parameters), MedGemma-4B-it (4B, deployed locally), and AntAngelMed (100B total, 6B active, MoE). All models are accessed via API except MedGemma-4B-it, which runs locally on 2×RTX 4090 GPUs.

B.2 Treatment-node Prediction Setting

All models are evaluated under the same treatment-node prediction setting. Given the available patient trajectory before each treatment adjustment, models are asked to predict three groups of outcomes: diagnostic capability, therapeutic efficacy, and medication safety.

For each treatment node, the input includes the patient information available before the target prediction point, including demographic information, clinical trajectory, physical examination, diagnostic tests, physician orders, and previous treatment history. The target labels are constructed from the corresponding effective observation window. Models do not receive future information beyond the prediction point.

The expected model output is a structured response containing the following components:

- **Diagnostic capability**: diagnosis, symptom extraction, and symptom severity;
- **Therapeutic efficacy**: Severity of Illness (SI), Global Improvement (GI), and Efficacy Index (EI);
- **Medication safety**: ADE occurrence, ADE extraction, ADE severity, suspected drug, and causality judgment.

B.3 Evaluation Metrics

Following the metric definitions in Section 5, we report task-specific primary metrics rather than using a single generic accuracy score for all tasks. All results are reported in percentage points.

For entity-level tasks, including diagnosis, symptom extraction, ADE extraction, and suspected-drug identification, we report F1 scores after deterministic normalization and strict matching. For diagnosis and ADE entities, the gold labels are expert-normalized to MedDRA terminology, and models are explicitly instructed to output standardized clinical terms. In the main evaluation, a predicted entity is counted as correct only if it exactly matches the corresponding MedDRA-normalized gold label after deterministic text normalization. For suspected drugs, predictions are matched against standardized drug names and, when applicable, normalized active ingredients.

For ordinal clinical scales, including symptom severity, SI, GI, EI, ADE severity, and causality, we report exact agreement with the annotated label as the primary metric. These tasks are treated separately from entity extraction because their labels correspond to predefined clinical grading scales rather than open-ended entity sets.

For ADE occurrence, we use binary event-detection accuracy as the primary metric in the main comparison. Additional diagnostic analyses, such as precision–recall behavior, terminology-level matching, and LLM-judged semantic equivalence, are reported separately in Appendix C.5. These

Table 9: Human reference performance on PedMed-Bench. The evaluator is a medical clinician with multiple years of clinical diagnostic experience. Results are reported using the same task-specific metrics as the main model comparison.

Diagnostic Capability and Efficacy						
Metric	**Diag.**	**Symp. Ext.**	**Symp. Sev.**	**SI**	**GI**	**EI**
Score	77.65	67.21	78.60	64.37	84.36	100.00

supplementary analyses are used to characterize error sources and terminology mismatch, but are not used as the primary scoring criterion in Table 2.

Aggregation Method. For metrics computed at the treatment-plan level (ADE severity and causality), we use patient-level macro-averaging: we first aggregate counts within each patient across all their treatment plans, then compute the metric, and finally average across patients. This patient-centric aggregation ensures that patients with more treatment plans do not disproportionately influence the overall score, aligning the evaluation with clinical practice where outcomes are ultimately assessed per patient rather than per intervention. For patients without valid predictions (e.g., no true positive ADEs detected), the corresponding metric is treated as missing for that patient and excluded from the average, as noted by "–" in Table 2.

B.4 Human Reference Performance

To contextualize the difficulty of PedMed-Bench, we report human reference performance for diagnostic capability and therapeutic efficacy in Table 9. The human evaluator is a medical clinician with multiple years of clinical diagnostic experience. We report this result as a reference rather than as an upper bound, since diagnostic reasoning and efficacy assessment under longitudinal pediatric trajectories still depend on case context, documentation quality, and protocol-specific label definitions. Therefore, the human reference should be interpreted as a contextual difficulty indicator for diagnosis and efficacy tasks, rather than a definitive clinical upper bound.

B.5 Prompting Strategy

Two-Phase Prompting Framework We have decomposed the pediatric clinical evaluation task into two sequential and logically linked phases, designed to simulate the iterative clinical reasoning of physicians moving from "status assessment" to "intervention feedback". This framework utilizes structured instructions to guide Large Language Models (LLMs) through evidence-based reasoning within complex clinical contexts, ensuring that the models comprehend the dynamic progression of the illness rather than focusing on a single point in time.

Phase 1: Diagnostic and Severity Baseline The core focus of the first phase is to establish an accurate clinical baseline. The primary component is Information Integration, where instructions guide the model to synthesize static diagnostic data with dynamic symptomatic manifestations to construct a comprehensive patient profile. The second focus is Baseline Anchoring; by introducing the Clinical Global Impressions-Severity (CGI-S) scale, we provide a standardized benchmark for the subsequent evaluation of therapeutic efficacy (Global Improvement) in the second phase. Finally, we implement Risk Calibration. Given the "high-risk management" environment typical of pediatric care, we explicitly incorporate "cautious and conservative" evaluation biases within the prompts to align the model's risk assessment logic with real-world pediatric clinical decision-making.

Phase 2: Interventional Efficacy and Safety Assessment The second phase focuses on the comprehensive evaluation following pharmacological intervention. Its primary objective is Temporal Logic, requiring the model to perform a deep comparison across the "past" and "present" timelines to precisely distinguish between pre-existing symptoms and new-onset ADEs. Regarding evaluation dimensions, this phase achieves Multi-scale Synergy by integrating the CGI-GI (Global Improvement), EI (Efficacy Index), CTCAE (Severity), and WHO-UMC/RUCAM (Causality Assessment) into a single prompt. Furthermore, to address semantic ambiguity in medical texts, we enforce strict Terminology Constraints. Through "negative constraints," the use of vague terms such as "allergy" is prohibited, forcing the model to output clinical manifestations with clear physical significance (e.g.,

"drug rash," "dyspnea"), thereby facilitating subsequent MedDRA standardization. Detailed prompt examples are provided below.

Phase 1: Baseline Assessment Prompt

[Role and Risk Calibration]
As a pediatric expert... please determine the patient's current disease diagnosis, symptoms, and the overall severity of illness. *(Risk Calibration Logic)*: Since the patient records are from China, the assessment of illness severity tends to be as thorough and cautious as possible. Predictions should lean toward a more cautious approach.

[Metric Integration and CGI-S Scale]
Choose severity from the CGI-S scale (options 1-7): (1) Normal, activities as usual; (4) Moderately ill, activity limited; (7) Extremely ill, respiratory failure.

[Output Format]
Output strictly in YAML format. Ensure prediction results are in Chinese but keep field names in English.

Phase 2: Treatment Feedback and Safety Reasoning Prompt

[Temporal Logic and ADE Filtering]
(Targeting Sequence Logic): If an event has already been recorded in the medical history, it should NOT be labeled as an adverse drug event, as it has already occurred.

[Multi-Scale Synergy: GI, EI, CTCAE, and Causality]
1. Global Improvement (CGI-GI): Select from (1) Very much improved to (7) Very much worse.
2. Severity and Causality: Refer to CTCAE v6.0 and WHO-UMC system. If the ADE is DILI, use RUCAM scale (score -9 to 14).

[Terminology Normalization Constraints]
(Targeting Normalization): The use of terms like "allergic reaction" is strictly prohibited. You must use specific clinical manifestations (e.g., drug rash, anaphylactic shock).

[Efficacy Index Calculation]
Notably, the EI (Efficacy Index) is derived as the quotient of two distinct metrics: the therapeutic efficacy assessment and the side effect evaluation.

B.6 Compute Resources

All experiments are inference-only evaluations; no model training or fine-tuning is performed. For hosted or API-based models, including proprietary frontier models and externally served open models, the provider-side hardware, memory, and execution environment are not directly observable. We therefore report the model identities, prompts, parsing rules, benchmark data, and evaluation code to support reproducibility.

For the locally executed medical-domain model, MedGemma-4B-it, inference was performed on a local workstation equipped with two NVIDIA RTX 4090 GPUs. The evaluation pipeline itself, including output parsing and metric computation, can be run on standard CPU resources after model outputs are obtained. Wall-clock time may vary depending on API latency, batching strategy, and local hardware configuration.

Table 10: Performance breakdown across pediatric developmental stages. Green cells indicate the best score within each metric, while red numbers indicate the lowest score. Results are reported in percentage points.

Metric		Infancy (>28 days to 1 year)	Early Childhood (>1–3 years)	Preschool Age (>3–6 years)	School Age (>6–12 years)	Adolescence (>12–18 years)
Diagnostic Capability	Diag.	**52.67**	50.26	**43.56**	44.28	46.88
	Symp. Ext.	23.48	**29.69**	23.46	28.72	**21.76**
	Symp. Sev.	43.91	46.78	45.61	**47.18**	**37.90**
Efficacy	SI	20.46	17.26	**15.23**	16.04	**20.86**
	GI	**33.05**	23.00	18.76	22.30	**14.05**
	EI	**26.22**	19.52	**16.52**	21.60	16.97
Safety	ADE Occ.	30.48	**36.21**	32.70	**24.30**	30.16
	ADE Ext.	1.14	2.18	**2.19**	0.58	**0.04**
	ADE Sev.	2.17	3.50	**3.91**	1.23	**0.13**
	Susp. Drug	0.23	**1.01**	0.69	0.19	**0.00**
	Causality	0.16	**1.33**	1.26	0.46	**0.00**

C Additional Analyses

Pediatric-Specific Failure Mode Analysis. The following subsections examine model performance through the lens of pediatric-specific characteristics that differentiate pediatric care from adult medicine. We stratify results by developmental stage (age groups), multidisciplinary consultation complexity, and concurrent medication count (polypharmacy burden). These analyses aim to uncover whether LLMs fail uniformly across pediatric patients or exhibit systematic weaknesses tied to age-dependent physiology, collaborative care complexity, and therapeutic intensity—factors that are central to pediatric medication safety and efficacy evaluation.

C.1 Performance across Pediatric Developmental Stages

Model performance varies non-linearly across pediatric developmental stages.

Pediatric medication reasoning is closely related to developmental stage, since physiology, pharmacokinetics, dosage thresholds, and contraindications change substantially from infancy to adolescence. We therefore stratify PedMed-Bench by five pediatric age groups: infancy, early childhood, preschool age, school age, and adolescence.

As shown in Table 10, model performance does not change monotonically with age. Diagnostic performance is highest in infancy but drops in the preschool group, while symptom extraction is strongest in early childhood and weakest in adolescence. Medication safety metrics also show clear age-dependent variation. ADE extraction, ADE severity, suspected-drug identification, and causality are relatively higher in early childhood or preschool-age groups, but approach the performance floor in adolescents. These results suggest that LLMs do not fail uniformly across pediatric patients; instead, different developmental stages expose different weaknesses in diagnosis, efficacy estimation, and medication safety attribution.

We interpret these results as descriptive subgroup findings rather than causal claims. The observed age-dependent gaps may reflect developmental pharmacology, disease distribution, medication complexity, documentation patterns, or sample-size variation within each age group.

C.2 Stratified Analysis by Clinical Complexity

C.2.1 By Number of Medications

Model performance becomes less stable as medication complexity increases. Figure 6 shows performance trends across different medication-count strata. As the number of medications increases, diagnosis and symptom-severity related metrics generally decline, suggesting that polypharmacy increases reasoning difficulty. In contrast, some safety-related metrics show modest gains in lower-complexity ranges, possibly because additional medications provide more explicit safety cues. However, performance fluctuations become more pronounced in high-medication cases, indicating that current LLMs remain unstable under complex multi-drug scenarios.

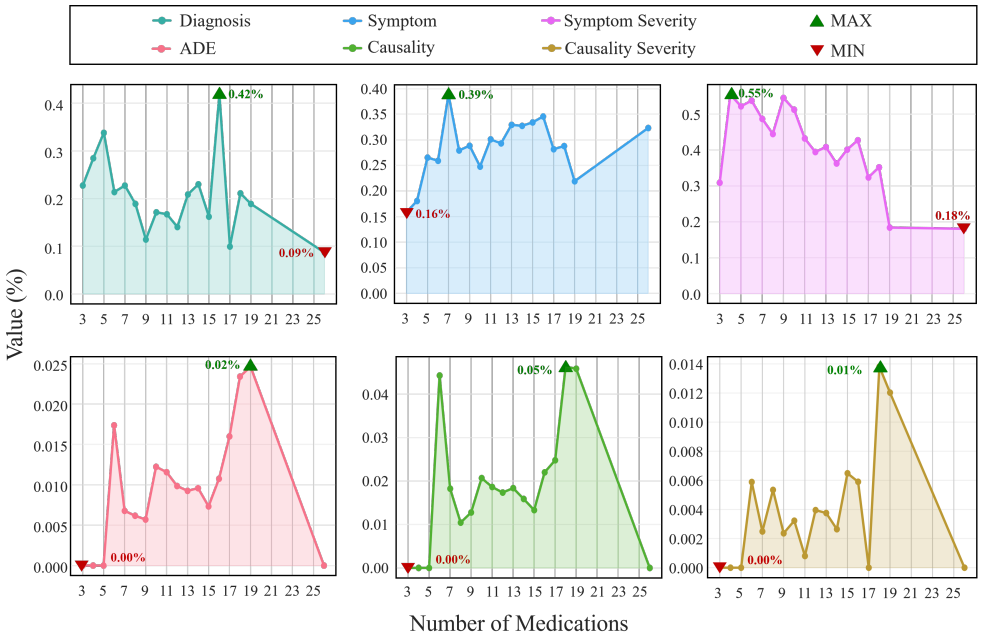


Figure 6: Performance stratified by medication count. As therapeutic complexity increases, models show greater performance instability across diagnostic, efficacy, and safety metrics, indicating that polypharmacy introduces additional reasoning challenges in pediatric medication evaluation.

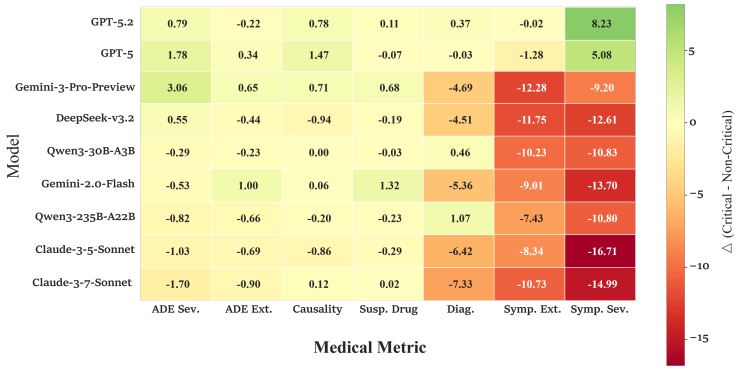


Figure 7: Performance delta between critical and non-critical cases. Negative values indicate lower performance on critical cases. Diagnosis and symptom-related metrics generally decline under critical conditions, while some safety metrics show small positive deltas, suggesting increased reasoning instability in high-acuity pediatric cases.

C.2.2 By Critical Status

Most models degrade on critical cases, indicating limited robustness in high-stakes settings. We compare model performance between non-critical and critical cases. As shown in Figure 7, most models exhibit lower performance on diagnosis, symptom extraction, and symptom severity in critical cases. This suggests that the increased urgency, heterogeneity, and documentation complexity of severe cases challenge reasoning stability. In contrast, some safety-related metrics show small improvements, possibly because critical cases contain more explicit treatment and adverse-event evidence. Overall, these results indicate that model robustness remains limited in high-stakes pediatric settings.

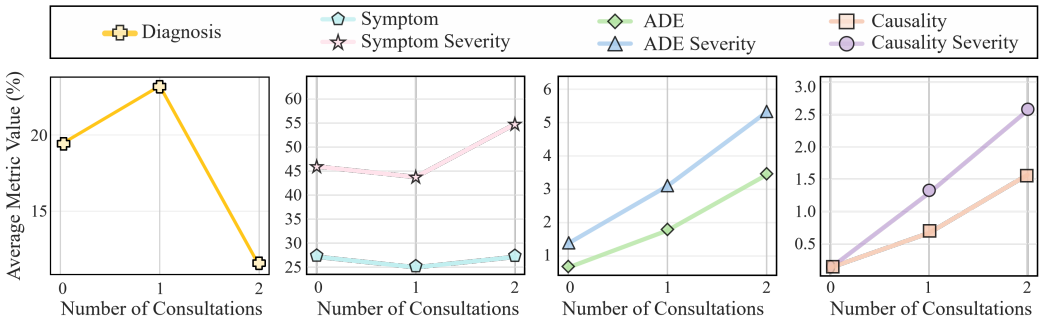


Figure 8: Performance stratified by the number of multidisciplinary consultations. Cases with consultation records show different diagnostic and safety prediction patterns, suggesting that consultation count reflects both additional clinical evidence and increased case complexity.

Table 11: Paired bootstrap comparison between *full* and *no_consult* settings. Confidence intervals are 95% bootstrap intervals.

Model	Metric	Δ	95% CI	p-value
GPT-5.4	ADE occurrence correctness	+0.0140	[-0.0335, +0.0559]	0.620
	Drug F1	-0.0007	[-0.0165, +0.0154]	0.953
	Safety tuple F1	+0.0011	[+0.0000, +0.0032]	0.265
DeepSeek-V3	ADE occurrence correctness	-0.0112	[-0.0391, +0.0196]	0.541
	Drug F1	+0.0030	[-0.0114, +0.0177]	0.737
	Safety tuple F1	+0.0000	[+0.0000, +0.0000]	1.000

C.3 Consultation-related Analysis

C.3.1 Stratified Analysis by Number of Consultations

Consultation count is associated with changes in model performance, but should be interpreted as a complexity marker rather than a direct causal factor. As shown in Figure 8, cases with consultation records exhibit different performance profiles from cases without consultation. Diagnostic and safety-related metrics tend to improve when one or more consultations are present, whereas symptom extraction and symptom severity show mild declines. This pattern suggests that consultation cases may provide richer expert evidence for higher-level reasoning, while also involving greater clinical ambiguity and documentation complexity. Therefore, the observed trend should be interpreted as a descriptive association rather than evidence that consultations directly improve model reasoning.

C.3.2 Consultation-note Ablation

Consultation notes are not a statistically significant direct driver of performance. To test whether consultation notes themselves materially improve model predictions, we conduct an ablation study on 358 cases with consultation records under three settings: *full* (complete EHR), *no_consult* (consultation notes removed), and *consult_only* (consultation notes only). Table 11 reports the results for GPT-5.4 and DeepSeek-V3.

For GPT-5.4, removing consultation notes changes ADE occurrence F1 from 0.4136 to 0.3958, drug F1 from 0.0864 to 0.0871, and safety tuple F1 from 0.0019 to 0.0008. For DeepSeek-V3, the corresponding changes are from 0.4375 to 0.4444, from 0.0810 to 0.0780, and from 0.0009 to 0.0009. Paired bootstrap tests show that the differences between *full* and *no_consult* are not statistically significant for either model. The *consult_only* condition also does not outperform the full EHR setting.

These results suggest that consultation count should be interpreted primarily as a marker of case complexity and documentation richness, rather than as direct evidence that consultation notes themselves improve model reasoning.

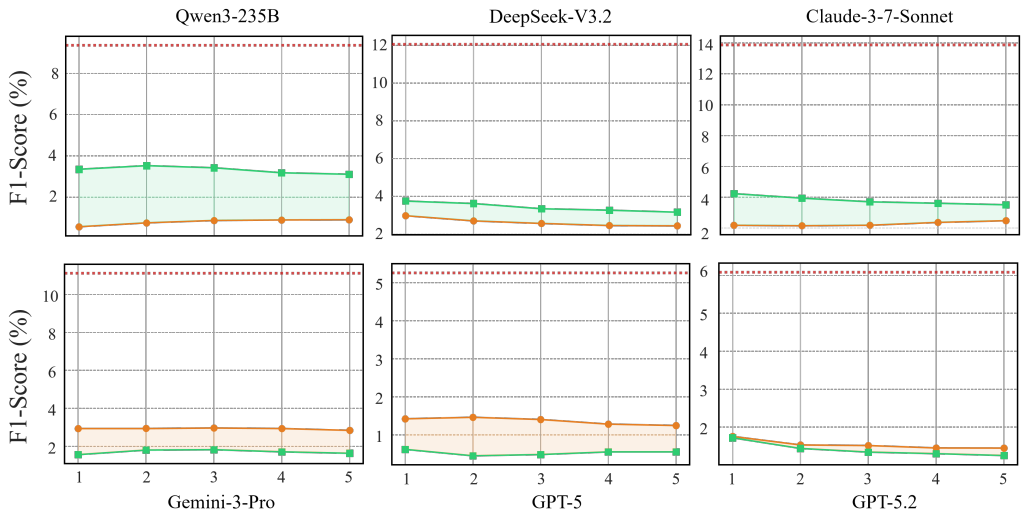


Figure 9: Temporal bias patterns in ADE prediction. The figure compares early and late ADE prediction tendencies across representative models, illustrating differences in how models respond to temporal evidence during medication safety assessment.

C.4 Temporal Bias Patterns in Safety Prediction

Different models exhibit distinct temporal bias patterns in ADE prediction. As shown in Figure 9, some models tend to produce earlier ADE-positive predictions, whereas others appear more conservative and respond only after stronger downstream evidence becomes available. For example, GPT-family models show a relatively earlier positive tendency in some settings, while other models exhibit a more delayed confirmation pattern. These results suggest heterogeneity in how models trade off sensitivity and evidential conservatism, which may be relevant for downstream deployment scenarios such as early screening versus confirmatory decision support.

C.5 Over-prediction and Terminology Matching

Over-prediction and standardized-label mismatch represent two distinct diagnostic failure modes. We further analyze diagnostic errors from two perspectives: prediction behavior and label alignment. First, as shown in Figure 3(a), models often achieve higher recall than precision in diagnosis and symptom extraction. This indicates an over-prediction tendency, where models generate broad candidate lists rather than precise clinical labels. While such behavior may be useful for early screening, excessive false positives can introduce diagnostic noise, unnecessary follow-up tests, and inappropriate treatment considerations in clinical workflows.

Second, we examine whether diagnostic errors arise from true clinical disagreement or from mismatch between free-form model outputs and standardized gold labels. In PedMed-Bench, gold diagnosis labels are normalized to MedDRA terminology, whereas model outputs are standardized text. We therefore evaluate diagnosis predictions under four matching levels: **Exact**, **LLT**, **PT**, and **Semantic**. Exact matching requires the model output to strictly match the MedDRA-normalized gold label. LLT-level and PT-level matching compare predictions and gold labels after normalization to the corresponding MedDRA levels, allowing increasingly coarser ontology-aware alignment. Semantic matching uses GPT-4o-mini OpenAI (2024) to assess whether the predicted diagnosis and gold diagnosis are medically equivalent, even if their standardized labels do not exactly match.

As shown in Figure 3(b) and Table 14, scores increase from Exact to LLT/PT and further under Semantic matching. This suggests that some free-form predictions are medically equivalent or clinically close to the gold diagnosis but fail strict standardized-label matching. However, Semantic matching should be interpreted cautiously: medical relatedness does not always guarantee sufficient specificity for standardized EHR coding, pharmacovigilance, handoff communication, or downstream decision support. Therefore, clinically useful pediatric models must improve not only diagnostic

Table 12: Human–LLM alignment for semantic diagnosis matching. Human semantic-equivalence judgments on 363 randomly sampled diagnosis prediction–gold label pairs are used as reference.

LLM Evaluator	Accuracy	Precision	Recall	F1
GPT-4o-mini	71.52	80.27	86.76	83.39

Table 13: Cross-judge consistency for semantic diagnosis matching. Diagnosis denotes strict exact matching against MedDRA-normalized gold labels, while the three semantic matching columns report supplementary LLM-judged medical equivalence scores. Scores are reported in percentage points.

Model	Diagnosis	Semantic Matching		
		GPT-4o-mini	GPT-5.4	Claude-4-6-Sonnet
DeepSeek-V3.2	19.42	46.23	35.21	49.17
Gemini-3-Pro-Preview	22.69	43.37	35.86	44.91
GPT-5	15.19	41.12	35.40	43.52
GPT-5.2	14.99	43.47	36.91	44.29
Qwen3-235B-A22B-Instruct	17.22	49.71	34.93	49.86
Qwen3-30B-A3B-Instruct	12.36	43.22	32.02	47.43

sensitivity, but also precision and standardized-label specificity. To validate the reliability of LLM-judged semantic matching, we further conduct human-alignment and cross-judge consistency analyses in Appendix C.6.

C.6 Validation of LLM-judged Semantic Matching

LLM-judged semantic matching is used only for supplementary error analysis. In the main evaluation, diagnosis predictions are scored by strict exact matching against MedDRA-normalized gold labels. Semantic matching is not used as the primary scoring criterion. Instead, it is introduced as a supplementary analysis to estimate how often strict label mismatch may arise from clinically equivalent but non-identical expressions.

To assess the reliability of the semantic matching judge, we randomly sampled 363 diagnosis prediction–gold label pairs and invited three medical master's students with clinical experience, all of whom had passed the annotation qualification test, to conduct blinded semantic-equivalence evaluation. Using the human judgment as reference, GPT-4o-mini achieved 71.52% accuracy and 83.39% F1, suggesting that the LLM judge is reasonably aligned with human clinical judgment for this supplementary analysis.

We further evaluate whether semantic matching results are specific to the GPT model family. To this end, we compare GPT-4o-mini with two alternative judge models, GPT-5.4 and Claude-4-6-Sonnet. As shown in Table 13, the semantic matching trends are broadly consistent across judge models. Claude-4-6-Sonnet, which is outside the GPT model family, produces similar or higher semantic matching scores in most cases. This suggests that the observed standardized-label mismatch is not solely an artifact of using a GPT-family judge.

Interestingly, GPT-4o-mini yields higher semantic matching scores than GPT-5.4 in this judging task. Manual spot checks suggest that stronger generative models may sometimes introduce redundant explanatory context when judging or aligning clinical expressions, which can reduce consistency under strict pairwise equivalence criteria. In contrast, GPT-4o-mini appears more directive and specific for this constrained semantic-equivalence task. We therefore use GPT-4o-mini as the default semantic judge in the supplementary analysis, while emphasizing that semantic matching is not used in the main benchmark score.

Table 14: Diagnosis matching under different label-alignment levels. Exact matching requires strict agreement with MedDRA-normalized gold labels. LLT and PT matching compare predictions after normalization to MedDRA LLT and PT levels. Semantic matching uses an LLM judge to assess medical equivalence between prediction and gold diagnosis. Scores are reported in percentage points.

Model	Exact F1	LLT F1	PT F1	Semantic F1
claude-4-6-sonnet	22.13	24.49	34.30	53.19
deepseek-v3.2	22.22	25.79	37.36	53.19
gemini-3-pro-preview	28.32	32.08	39.81	54.13
gpt-5	18.67	10.32	13.48	51.36
gpt-5.2	19.39	10.67	13.72	55.94
gpt-5.4	19.17	23.53	31.69	51.91
qwen3-235b-a22b-instruct-2507	19.27	23.63	34.62	56.89
qwen3-30b-a3b-instruct-2507	14.33	20.22	27.04	50.44

D Language Bias Analysis

To ensure that our benchmark results reflect clinical reasoning capability rather than language-specific artifacts, we conducted two complementary analyses:

Cross-lingual Medical QA Validation. We evaluated six representative models on CNMLEQA-3k Zong et al. (2026), a Chinese medical licensing exam benchmark comprising 600 questions. This external validation serves two purposes: (1) it confirms that models retain strong medical knowledge when processing Chinese clinical text, and (2) it establishes that our translation from Chinese EHR data to English did not introduce systematic distortions. Results show that all evaluated models achieve end-to-end accuracy above 78%, with Gemini-3-Pro-Preview reaching 84.0% (Table 18).

Paired Language-Bias Benchmark. We constructed a controlled 121-patient subset (150 treatment records) where each clinical case was presented in both Chinese (original) and human-verified English (translated). Six models were evaluated on identical cases in both languages under the same evaluation protocol. We applied paired Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction across 36 model-by-metric comparisons.

Among 36 comparisons, only 2 reached FDR-adjusted significance: Claude-4.6-Sonnet showed higher Symptom F1 on Chinese input (+4.2 percentage points, FDR $p = 0.0022$), while GPT-5.2 favored English (+9.5 percentage points, FDR $p = 0.0002$). For the primary diagnostic capability metric (Diagnosis overall F1), no model exhibited significant language bias (all FDR $p > 0.41$). The mean absolute difference across all models and metrics was 1.8 percentage points, indicating that language choice introduces minimal systematic bias in PedMed-Bench evaluation (Table 19).

These findings support our decision to translate the original Chinese EHR data into English: the translated benchmark preserves clinical reasoning demands while enabling broader international participation and comparison with English-dominant frontier models.

Table 15: PedMed-Bench annotation metrics, reference standards, and scoring levels.

Abbr.	Reference Standard	Grade / Score	Detailed Description and Clinical Indicators
Diag.	Clinical Gold Standard	LLT / PT	Diagnosis mapped to MedDRA Lowest Level Term and Preferred Term.
Symp. Ext.	Clinical Records	Set of entities	Patient symptoms and signs extracted from clinical notes.
Symp. Sev.	Clinical Guidelines	0–2	0: Mild; 1: Moderate; 2: Severe functional impact.
SI	CGI Scale	1: Normal	No cough or wheezing; behaves like a healthy child.
		2: Very Mild	Occasional cough; no fever; no daily-life impact.
		3: Mild	Daily cough or minor wheezing; attends school; no oxygen.
		4: Moderate	Restricted activity or nocturnal awakening.
		5: Mod. Severe	Obvious symptoms; irritable or lethargic; requires oxygen.
		6: Severe	Dyspnea or cyanosis; requires ICU monitoring.
		7: Ext. Severe	Respiratory failure, mechanical ventilation, or unstable signs.
GI	CGI Scale	1: Very Much Imp.	Symptoms nearly resolved; care level downgraded.
		2: Much Imp.	Major symptoms reduced $> 50\%$; medication can be reduced.
		3: Minimally Imp.	Slight symptom relief compared with previous course.
		4: No Change	No observable clinical change.
		5: Minimally Worse	Slight worsening without new symptoms.
		6: Much Worse	Worsened symptoms with new diagnoses or added medication.
		7: Very Much Worse	Cyanosis, respiratory failure, or urgent intervention required.
EI	CGI Scale	0–3 (Eff.)	3: Marked; 2: Effective; 1: Minimal; 0: No change or worse.
		0–3 (Side)	0: None; 1: Mild; 2: Moderate; 3: Severe safety impact.
ADE Occ.	Clinical Records	0 / 1	0: No ADE; 1: ADE absent at admission but emerging in the observation window.
ADE Ext.	MedDRA / Labeling	LLT / PT	ADE name mapped to MedDRA LLT and PT.
ADE Sev.	CTCAE v6.0	Grade 1–5	G1: Mild; G2: Moderate; G3: Severe; G4: Life-threatening; G5: Death.
	Hy's Law	0: Mild, 1: Severe	Severe Drug-Induced Liver Injury (DILI) if ALT/AST $> 3\times$ULN and bilirubin $> 2\times$ULN.
Susp. Drug	Drug Labeling	Set of entities	Suspected drugs verified against labeling and patient history.
Causality	WHO-UMC	1–6	6: Certain; 5: Probable; 4: Possible; 3: Unlikely; 2: Conditional; 1: Unassessable.
	RUCAM	1: Excluded	Score ≤ 0.
		2: Unlikely	Score 1–2.
		3: Possible	Score 3–5.
		4: Probable	Score 6–8.
		5: Highly Prob.	Score ≥ 9; total range: -9 to 14.

Table 16: Distribution of Clinical Diagnoses by MedDRA System Organ Class (SOC), pooled across both centers ($N = 316$)

SOC Name	LLT Name (Frequency)	Total Freq.	Comp. (%)
Infections and infestations	Bronchial pneumonia(105), Acute upper respiratory tract infection(53), Acute bronchitis(42), Community acquired pneumonia(29), Infectious pneumonitis(28), Infectious mononucleosis(22), Pneumonia aggravated(19), Epstein-Barr virus infection(14), Acute gastroenteritis(12), Mycoplasma infection(10), EBV infection(6), Pyemia(5), Acute laryngitis(5), 2019 novel coronavirus infection(5), Pneumonia NOS(4), Chronic sinusitis(3), Acute rhinitis(3), Otitis media acute(3), Bronchitis(3), Infection upper respiratory(3), Oral candida albicans infection(2), Adenovirus infection(2), Acute myringitis(2), Viral exanthem(2), Acute hepatitis A(2), Infective myositis(2), Acute tracheobronchitis(2), Secretory otitis media(2), Tonsillitis pyogenic(1), Lobar pneumonia(1), Chronic urinary tract infection(1), Acute sinusitis(1), Skin infection(1), BK virus infection(1), Severe fever with thrombocytopenia syndrome(1), Otitis media(1), Toxoplasmosis(1), Lung infection(1), Aspiration pneumonia(1), Neonatal infection NOS(1), Cytomegalovirus infection(1), Sinusitis(1), Chlamydial infection(1), Ethmoidal sinusitis(1)	406	39.57%
Hepatobiliary disorders	Liver damage(125), Drug-induced liver disease(33), Fatty liver(27), Hyperbilirubinaemia neonatal(1), Damage liver(1), Function liver abnormal(1)	188	18.32%
Metabolism and nutrition disorders	Hypokalaemia(17), Diabetes(9), Acidosis metabolic(6), Adipositas(6), Dehydration(5), Hyperlactacidaemia(5), Diabetic ketoacidosis(4), Hyponatraemia(2), Dietary deficiency(2), Hypoglycaemia(2), Type 2 diabetes mellitus(2), Obesity(2), Hyperkalaemia(2), Neonatal hypocalcaemia(2), Hypocalcaemia(2), Hypovitaminosis D(1), Metabolic syndrome(1), Late metabolic acidosis of newborn(1), Asymptomatic hyperlactataemia(1), Hypoproteinaemia(1), Hyperuricaemia(1), Neonatal hyponatraemia(1), Hypoglycemia(1)	76	7.41%
Blood and lymphatic system disorders	Neutropenia(24), Anaemia(10), Anaemia haemolytic DCN(10), Clotting disorder(4), Mesenteric adenitis(4), Lymphadenitis cervical(3), Thrombocytopaenia(2), Agranulocytosis(2), Haemolytic autoimmune anaemia DCN(2), Hypersplenotrophy(2), Anaemia aplastic aregenerative(1), Cervical adenitis(1), Lymph nodes cervical swollen(1), Lymphocytosis(1), Primary thrombocytopenia(1), Idiopathic thrombocytopenia(1), Anaemia neonatal(1)	70	6.82%

Table 17: Distribution of Adverse Drug Events (ADEs) by MedDRA System Organ Class (SOC), pooled across both centers ($N = 316$)

SOC Name	LLT Name (Frequency)	Total Freq.	Comp. (%)
Hepatobiliary disorders	Damage liver(117)	117	26.96%
Gastrointestinal disorders	Diarrhea(38), Nausea and vomiting(31), Abd. pain(11), Loose bowel(8), Constipation(6), Emesis(5), Abdo. discomfort(3), Abdominal distension(3), Dry stools(2), Intussusception(1), Possetting(1), Infantile diarrhoea(1), Secretory diarrhea(1), Functional diarrhea(1), Nausea(1)	113	26.04%
Skin and subcutaneous tissue disorders	Exanthema(43), Papular rash(21), Exanthema generalised(9), Dermatitis medicamentosa(2), Livedo reticularis(1), Yellow skin(1), SLE rash(1), Papular rash on hands(1)	79	18.20%
Cardiac disorders	Myocardial injury(29), Acute myocardial ischaemia(1), Cardiac dysrhythmias(1)	31	7.14%
Investigations	Neutrophils reduced(9), Creatine kinase increased(3), White blood cell decreased(3), Cardiac enzymes increased(1), WBC dec(1), Platelets decreased(1), Blood bilirubin increased(1)	19	4.38%
Blood and lymphatic system disorders	Anaemia(10), Agranulocytosis(3), Haemolytic autoimmune anaemia DCN(2), Clotting disorder(1), Leukocytosis(1), Neutropenia(1)	18	4.15%
Eye disorders	Eye discomfort(8), Blepharal edema(3)	11	2.53%
Respiratory, thoracic and mediastinal disorders	Bleeding nose(4), Adenoidal hypertrophy(1), Nasal obstruction(1), Acidosis respiratory(1), Snore(1)	8	1.84%
Nervous system disorders	Infant irritability(4), Shaking of hands(3), Cephalalgia(1)	8	1.84%
Infections and infestations	Pyemia(3), Oral candida albicans infection(1), Chronic urinary tract infection(1), Skin infection(1)	6	1.38%
Musculoskeletal and connective tissue disorders	Muscle pain(2), Muscle twitching(1), Pain of lower extremities(1), Muscle ache(1)	5	1.15%
Metabolism and nutrition disorders	Hypokalaemia(3), Decreased appetite(1), Hyperkalaemia(1)	5	1.15%
Ear and labyrinth disorders	Ear ache(4)	4	0.92%
Immune system disorders	Drug allergy(3), Allergic reaction to drug(1)	4	0.92%
General disorders and administration site conditions	Chest tightness(2), Dysthermia(1)	3	0.69%
Renal and urinary disorders	Hematuria(1)	1	0.23%
Vascular disorders	Blue skin(1)	1	0.23%
Injury, poisoning and procedural complications	Muscular injury(1)	1	0.23%
Total		**434**	**100.00%**

Table 18: Performance on CNMLEQA-3k Chinese Medical Licensing Exam (600 questions). End-to-end accuracy measures the proportion of correctly answered questions among all 600; conditional accuracy reports accuracy on valid responses only (coverage $> 90\%$ for all models).

Model	Coverage	Valid	Cond. Acc.	E2E Acc.
Gemini-3-Pro-Preview	90.8%	545	92.5%	84.0%
Claude-4.6-Sonnet	93.5%	561	88.9%	83.2%
GPT-5.4	93.5%	561	88.4%	82.7%
GPT-5.2	93.2%	559	86.4%	80.5%
GPT-5	93.5%	561	85.7%	80.2%
DeepSeek-V3.2	93.5%	561	83.4%	78.0%

Table 19: Paired Chinese–English language-bias analysis on 121 patients (150 treatment records). Only comparisons with FDR-adjusted $p < 0.05$ are shown. The complete 36-comparison table is available in supplementary materials.

Model	Metric	CN mean	EN mean	EN−CN	95% CI
Claude-4.6-Sonnet	Symptom F1	0.334	0.292	-0.042	$[-0.067, -0.016]$
GPT-5.2	Symptom F1	0.144	0.239	$+0.095$	$[+0.057, +0.134]$