

Beyond Static EHRs: Dynamic Examination Feedback for LLM-based Virtual Patients

Renwei Xiao¹ Wen He⁴ Hongzhou Yu¹ Qingqiu Li¹ Yuejie Zhang¹
Rui Feng^{1,2,3,4*} Xiaobo Zhang^{4*}

¹College of Computer Science and Artificial Intelligence, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University ²Fudan Zhangjiang Institute

³College of Intelligent Robotics and Advanced Manufacturing, Fudan University

⁴Children's Hospital of Fudan University, National Children's Medical Center

{rxiao23,hzyu24,qqli22}@m.fudan.edu.cn, {hewen,yjzhang,fengrui}@fudan.edu.cn, zhangxiaobo0307@163.com

Abstract

Large language models (LLMs) have shown significant promise in healthcare applications. To better mirror real-world settings for LLM evaluation, dynamic longitudinal diagnosis-and-treatment simulation with virtual patients has recently emerged as a focal point of research. However, existing simulation frameworks are limited by clinical trajectory uniqueness, where virtual patients can only provide feedback based on information available in static electronic health records (EHRs). This limitation leads to simulation failures when unrecorded but medically sound examinations are ordered. In this paper, we formulate the task of Examination Feedback Simulation, which aims to dynamically augment the unique trajectory by simulating medically plausible examination results in response to clinical orders. To support this largely unexplored research area, we construct ClinTrack, a dataset curated from MIMIC-IV. We further propose a structure-aware evaluation metric SimScore to quantitatively assess the quality of simulated results, which shows promising initial alignment with expert judgment. Building on this framework, we develop ClinSim, a tailored generative model for this task. Experiments demonstrate that our 4B ClinSim model significantly outperforms flagship models with up to 235B parameters on this task, providing a critical foundation for creating more dynamic and realistic virtual patients. Code and other resources are available at [GitHub](https://github.com/X-Ray-git/clinsim)¹.

1 Introduction

Artificial intelligence (AI), particularly large language models (LLMs), is considered a promising solution to address the long-standing challenges in healthcare, such as access, quality, and cost containment (Kissick, 1994; Wang et al., 2024;

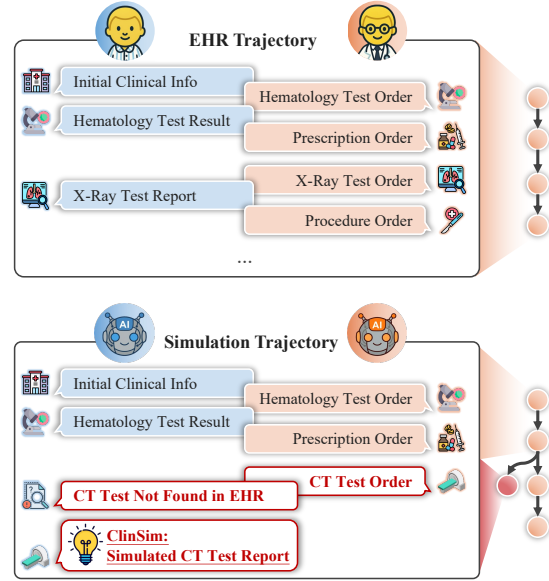


Figure 1: The limitation of clinical trajectory uniqueness in static EHRs. By referring to EHRs (top), the LLM-based virtual patient (bottom left) provides feedback to the LLM doctor (bottom right). It functions well until the doctor orders an unrecorded examination. Our approach aims to support examination feedback simulation beyond the recorded clinical trajectory.

OpenAI et al., 2024; Brodeur et al., 2025; Singhal et al., 2023). To provide a more realistic basis for evaluating LLM performance, dynamic longitudinal diagnosis-and-treatment simulation has recently emerged as a focal point of research (Luo et al., 2026; Nori et al., 2025; Zhou et al., 2025; Shang et al., 2025; Li et al., 2024; Fansi Tchango et al., 2022; Baichuan-M2 Team et al., 2025; Liu et al., 2025b; Johri et al., 2024; Liu et al., 2025a; Schmidgall et al., 2026). Conceptually paralleling the Objective Structured Clinical Examination (OSCE) widely utilized to assess human clinical competence (Harden et al., 1975; Tu et al., 2025), this simulation paradigm represents a crucial step towards mirroring real-world clinical practice. In these simulation frameworks, the medical LLM

*Corresponding author.

¹<https://github.com/X-Ray-git/clinsim>

agent acts as a doctor, and another LLM agent serves as a virtual patient that simulates the actions of sequentially reporting symptoms and examination feedback to the doctor by referring to electronic health records (EHRs). After the simulation, the performance of the medical LLM is evaluated based on the entire clinical interaction process with the virtual patient. Among various clinical actions, the medical examination (including radiological, microbiological, and laboratory examinations) is one of the most essential, as the ability to order and interpret it is a fundamental component of medical competence (Zayed et al., 2025; Amrollahi et al., 2024; Johnson et al., 2019).

However, the simulation of medical examinations in prior studies is inadequate. Existing simulation frameworks inherit the limitation of clinical trajectory uniqueness naturally present in static EHRs, implicitly excluding any potentially reasonable deviation from the recorded clinical actions. This limitation, illustrated in Figure 1, creates a critical failure point: as existing virtual patients can only provide examination feedback based on the unique and static EHRs (Zhou et al., 2025; Liu et al., 2025b), ordering any unrecorded examination, even if medically sound, can cause the interaction to halt (Nori et al., 2025; Shang et al., 2025). The virtual patient’s refusal to provide feedback on unrecorded examinations degrades the evaluation from a genuine assessment of clinical competence to a mere test of whether the LLM doctor can guess the specific examinations listed in the original record.

To address this limitation, we present a comprehensive, three-part solution. (1) We formulate the task of **Examination Feedback Simulation**, which involves leveraging a patient’s clinical history to simulate medically plausible examination results in response to medical orders. To support this task, we introduce ClinTrack-MIMIC (**ClinTrack** for simplicity), a dataset curated from MIMIC-IV (Johnson et al., 2023) that enables scalable, reference-anchored evaluation serving as a proxy for the intended off-path simulation setting. (2) We propose **SimScore**, an objective, rule-based evaluation metric that leverages the structured nature of our data to quantitatively assess the simulated results. A blind review by medical experts on a randomly selected sample set provides preliminary evidence that SimScore aligns with expert judgments, underscoring its potential as an objective metric for quantitative evaluation. (3) We develop **ClinSim**,

a specialized generative model, to demonstrate the potential of a purpose-built LLM in this task. ClinSim, with only 4 billion parameters, significantly outperforms various larger models (up to 235B, in few-shot settings).

Our key contributions can be summarized as follows:

- We systematically define the task of examination feedback simulation, and introduce the supporting ClinTrack dataset, laying the groundwork for creating more realistic simulation frameworks.
- We propose SimScore, a rule-based metric derived from a structure-aware evaluation schema, to assess the quality of the simulated results, and investigate its alignment with expert judgment in an initial human evaluation.
- We develop ClinSim, a model specifically trained for this task that achieves superior performance compared to existing models, indicating that the capability of simulating examination feedback is highly trainable, and demonstrating the feasibility of effectively instilling task-specific capabilities into existing LLMs.

2 Related Work

2.1 Virtual Patient

Clinical evaluation of LLMs has long relied on static datasets and isolated scenarios that fail to capture the cascading effects of healthcare decisions (Luo et al., 2026). To address this limitation, dynamic longitudinal diagnosis-and-treatment simulation with virtual patients has recently emerged as a focal point of research.

In multi-turn clinical dialogue simulation (Johri et al., 2024; Liu et al., 2025a; Sun et al., 2025; Liu et al., 2025b; Fansi Tchango et al., 2022; Li et al., 2024; Tu et al., 2025; Baichuan-M2 Team et al., 2025), the medical model incrementally collects information from a virtual patient role-played by another model. These approaches reveal that the diagnostic performance of the medical model declines significantly when it is required to gather information progressively through a multi-turn dialogue, as opposed to receiving the entire record at once to make a final diagnosis (Johri et al., 2024; Liu et al., 2025a). SD-Bench (Nori et al., 2025), DynamicCare (Shang et al., 2025), MultiCogEval (Zhou

et al., 2025) and AgentClinic (Schmidgall et al., 2026) propose frameworks that allow the LLM doctor to order clinical examinations and treatments, receiving feedback based on real patient records.

2.2 Examination Feedback Simulation

Examination Feedback Simulation is a latent component of these simulation frameworks. It aims to address the limitation of clinical trajectory uniqueness in static EHRs. Due to this limitation, any attempt to order a reasonable but unrecorded examination inevitably halts the interaction. Early research intentionally sidesteps this challenge by relaxing the requirement of “longitudinal”, excluding new examination orders from the simulation scope (Fansi Tchango et al., 2022; Liu et al., 2025b). In later studies, AgentClinic (Schmidgall et al., 2026) mitigates this issue by prompting the virtual patient to respond with “Normal Readings” for unrecorded examinations. MediQ (Li et al., 2024) and DynamiCare (Shang et al., 2025) instruct the virtual patient to faithfully respond based on the original records, leading to a possible refusal to provide corresponding feedback.

Nori et al. (2025) highlight a key weakness of this approach: refusals can imply the ground-truth trajectory and discourage valid alternative ones. To overcome this, they employ a language model to simulate feedback for unrecorded examinations. The simulation is treated as an auxiliary function rather than a primary research focus, and is not systematically and quantitatively validated. Luo et al. (2026) also propose a theoretical framework of the Clinical Environment Simulator (CES) for generating counterfactual trajectories, but it remains a high-level perspective rather than a practical tool.

3 ClinTrack Dataset Curation

Our study is based on MIMIC-IV (Johnson et al., 2023), a large-scale, publicly available dataset well-suited for constructing a dynamic longitudinal diagnosis-and-treatment simulation framework. MIMIC-IV is widely used for clinical research and is organized into multiple tables linked by unique patient and admission identifiers.

The final data structure is exemplified in Figure 2. To construct the ClinTrack dataset, we follow a 4-step process comprising 3 stages as shown in Figure 3.

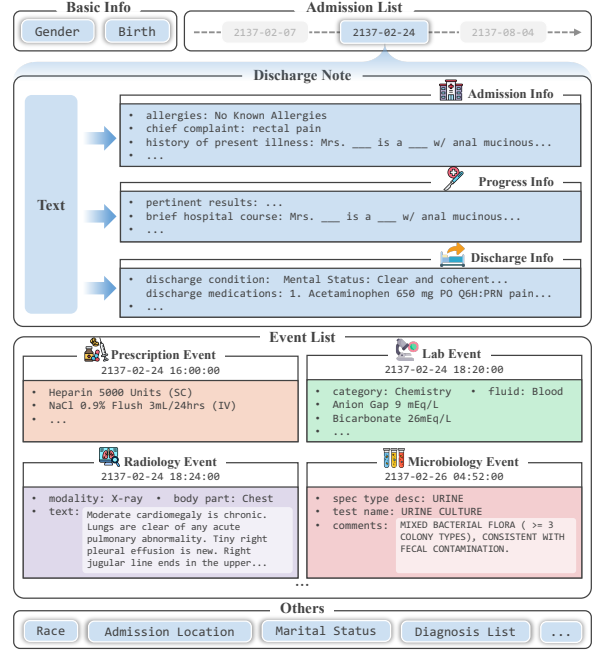


Figure 2: Data structure of a record in ClinTrack. Another detailed JSON example can be found in Appendix H.

3.1 Stage 1: Data Aggregation

The initial stage involves gathering information across all tables and reconstructing each patient’s clinical history into a chronologically ordered sequence of events.

Step ① Each record, denoted as R , is structured as a tuple containing basic patient information I_{basic} , an unstructured discharge note I_{note} , a list of diagnoses I_{diag} , and a sequence of clinical events $I_{\text{seq}} = \{E_i\}_{i=1}^N$, where E_i represents the i -th clinical event in chronological order.

$$R = (I_{\text{basic}}, I_{\text{note}}, I_{\text{diag}}, I_{\text{seq}}) \quad (1)$$

3.2 Stage 2: Fine-grained Parsing

The second stage focuses on fine-grained parsing of the clinical text to extract relevant information.

Step ② ③ To leverage the rich information contained in the unstructured notes, we employ an LLM to extract structured data. For discharge notes, this includes admission information I_{adm} , progress information I_{prog} , and discharge information I_{disch} . The detailed schema of these fields is provided in Appendix G. For radiology reports, we extract modality information I_{modality} and body part information $I_{\text{body_part}}$.

$$(I_{\text{adm}}, I_{\text{prog}}, I_{\text{disch}}) = \mathcal{LLM}(p_{\text{note_extract}}, I_{\text{note}}) \quad (2)$$

$$(I_{\text{modality}}, I_{\text{body_part}}) = \mathcal{LLM}(p_{\text{report_extract}}, I_{\text{report}}) \quad (3)$$

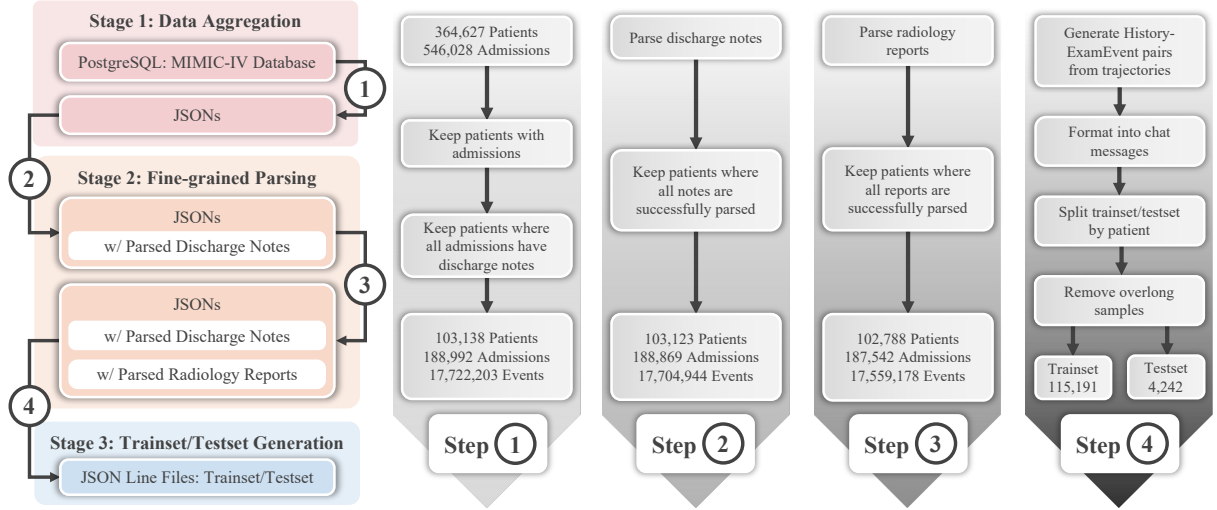


Figure 3: The dataset curation process consists of three stages implemented across four steps. The JSON schemas from the first three steps are structurally compatible, allowing records from Steps 2 and 3 to be seamlessly merged back into the dataset for augmentation. In this work, we proceed to Step 4 utilizing only the most comprehensively annotated records from Step 3.

where I_{report} is a field in E_i if the event type is *RadiologyEvent*, and p denotes the corresponding prompt used for the LLM.

3.3 Stage 3: Trainset and Testset Generation

The third stage involves transforming the structured data into prompts for training and evaluation purposes.

Step ④ We frame the core functionality of the expected model as a conditional generation task. Given a specific time step i , the preliminary information of the patient’s record is denoted as $S_{\text{pre}}^{(i)}$. To better specify the latent patient state in our retrospective simulation setting, we additionally incorporate the final diagnosis list I_{diag} as auxiliary patient-side case information. This information is available only to the virtual-patient simulator and is never exposed to the doctor agent. The model is tasked with generating the subsequent clinical event, E_i , conditioned on the preliminary information, the auxiliary final diagnosis list, and the type and time of that event. The event type is constrained to one of $\{\text{Radiology}, \text{Microbiology}, \text{Lab}\}$. Each event type can be further described by subtypes, which are also provided as part of the input to the model. For *Radiology* events, the subtype consists of the modality (such as *CT*, *X-Ray*, and *Ultrasound*) and the body part (such as *Chest*, *Upper Extremity*, and *Head*). For *Microbiology* events, the subtype is the type of culture (such as *BloodCulture*, *UrineCulture*, and *AcidFastCulture*), and the specimen type (such as *Stool*, *Blood*, and

Urine). For *Lab* events, the subtype is the type of fluid (such as *Blood*, *Urine*, and *JointFluid*), and the category (such as *Chemistry*, *Hematology*, and *BloodGas*). The task is formulated as follows:

$$\hat{E}_i = \mathcal{LLM}(p_{\text{sim}}, S_{\text{pre}}^{(i)}, I_{\text{diag}}, \text{Type}(E_i), \text{Time}(E_i)) \quad (4)$$

$$S_{\text{pre}}^{(i)} = (I_{\text{basic}}, I_{\text{adm}}, E_1, E_2, \dots, E_{i-1}) \quad (5)$$

where p_{sim} is the prompt used to instruct the model to perform the simulation task.

4 SimScore: Structure-Aware Evaluation

Existing evaluation metrics, whether based on text overlap (e.g., ROUGE (Lin, 2004), Levenshtein (Levenshtein, 1965)) or semantic embeddings (e.g., BERTScore (Zhang et al., 2020)), treat the output as plain text. Since our simulation produces structured clinical data, these traditional metrics are suboptimal, as they overlook the distinct semantics of specific data types (e.g., numerical values), intrinsic structures, and the clinical significance of different JSON fields. To address this, we propose a domain-specific metric, SimScore, to quantify the fidelity of the simulated clinical events.

Instead of using a generic recursive JSON comparison, SimScore $s = \mathcal{S}_{\text{type}}(E_g, \hat{E})$ is computed by routing the ground-truth event E_g and the predicted event $\hat{E}$ to the corresponding specific scorer based on the event type:

$$\mathcal{S}_{\text{type}}(E_g, \hat{E}) = \begin{cases} \mathcal{S}_{\text{rad}}(E_g, \hat{E}) & \text{if type} = \text{Radio} \\ \mathcal{S}_{\text{lab}}(E_g, \hat{E}) & \text{if type} = \text{Lab} \\ \mathcal{S}_{\text{micro}}(E_g, \hat{E}) & \text{if type} = \text{Micro} \end{cases} \quad (6)$$

where each scorer is composed of several primitive metrics.

4.1 Primitive Metrics

We define four foundational primitive metrics to compare primitive values.

Textual fields are evaluated using a composite similarity σ_{text} that averages Levenshtein similarity and ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) to capture both character-level robustness and phrase overlap.

$$\sigma_{\text{text}}(v_g, v_t) = \frac{1}{4} (\sigma_{\text{lev}}(v_g, v_t) + R_1(v_g, v_t) + R_2(v_g, v_t) + R_L(v_g, v_t)) \quad (7)$$

Numeric fields are evaluated based on the relative error.

$$\sigma_{\text{num}}(v_g, v_t) = \frac{1}{1 + \frac{|v_t - v_g|}{|v_g| + \epsilon}} \quad (8)$$

Categorical or identifier fields use an exact match indicator σ_{exact} .

$$\sigma_{\text{exact}}(v_g, v_t) = \mathbb{1}_{v_g = v_t} \quad (9)$$

Set-based fields use the standard Jaccard similarity.

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (10)$$

4.2 Radiology Scorer ($\mathcal{S}_{\text{rad}}$)

For radiology events, the primary clinical information is encapsulated in the report. Thus, the scorer exclusively evaluates the text field, ignoring other structural metadata.

$$\mathcal{S}_{\text{rad}}(E_g, \hat{E}) = \sigma_{\text{text}}(E_g[\text{text}], \hat{E}[\text{text}]) \quad (11)$$

4.3 Laboratory Scorer ($\mathcal{S}_{\text{lab}}$)

Laboratory events typically contain a list of multiple test items. We first match the predicted items with the ground truth based on their label. For each matched entry pair (e_g, e_t) , we compute a composite score $\sigma_{\text{lab_entry}}$ using a weighted sum of the scores for the value σ_{val} , abnormality flag, and unit of measurement (UOM). Let $e_g^{(n)}$ and $e_g^{(v)}$

denote the valuenum and value fields of a ground truth entry e_g , and $e_t^{(n)}$ and $e_t^{(v)}$ denote the corresponding fields of a predicted entry e_t . The score σ_{val} dynamically adapts to the data modality of the ground truth:

$$\sigma_{\text{val}}(e_g, e_t) = \begin{cases} \sigma_{\text{num}}(e_g^{(n)}, e_t^{(n)}) & \text{if } e_g^{(n)} \neq \text{null} \\ \sigma_{\text{text}}(e_g^{(v)}, e_t^{(v)}) & \text{otherwise} \end{cases} \quad (12)$$

The composite entry score is then formulated as follows, where we empirically set the weights to $w_v = 0.70$, $w_f = 0.15$, and $w_u = 0.15$ to reflect their relative clinical importance.

$$\begin{aligned} \sigma_{\text{lab_entry}}(e_g, e_t) &= w_v \cdot \sigma_{\text{val}}(e_g, e_t) \\ &+ w_f \cdot \sigma_{\text{exact}}(e_g[\text{flag}], e_t[\text{flag}]) \\ &+ w_u \cdot \sigma_{\text{exact}}(e_g[\text{valueuom}], e_t[\text{valueuom}]) \end{aligned} \quad (13)$$

The final score is computed by aggregating the entry scores and normalizing them by the size of the union of label sets $|L_g \cup L_t|$ across both records. This union-based normalization acts as a strict penalty mechanism, ensuring the final score is reduced for any omitted or hallucinated entries.

$$\mathcal{S}_{\text{lab}}(E_g, \hat{E}) = \frac{\sum \sigma_{\text{lab_entry}}(e_g, e_t)}{|L_g \cup L_t|} \quad (14)$$

4.4 Microbiology Scorer ($\mathcal{S}_{\text{micro}}$)

Microbiology events are evaluated based on the clinical outcome. If neither the ground truth nor the prediction contains pathogenic organisms ($\mathbb{O}_g = \emptyset \wedge \mathbb{O}_t = \emptyset$), the structural fields are inherently empty. In this scenario, the critical clinical context (e.g., “no growth” or specimen quality) is entirely encapsulated within the comments field C . Otherwise, we construct a set of atomic fact tuples $\mathbb{T}$ (where each tuple comprises the organism name, antibiotic name, and sensitivity interpretation) and compute the Jaccard similarity between the sets $\mathbb{T}_g$ and $\mathbb{T}_t$.

$$\mathcal{S}_{\text{micro}}(E_g, \hat{E}) = \begin{cases} \sigma_{\text{text}}(C_g, C_t) & \text{if } \mathbb{O}_g = \emptyset \wedge \mathbb{O}_t = \emptyset \\ \text{Jaccard}(\mathbb{T}_g, \mathbb{T}_t) & \text{otherwise} \end{cases} \quad (15)$$

5 Experiments

5.1 Experimental Setup

To rigorously prevent data leakage, we partition the ClinTrack dataset into a trainset and a testset at the patient level with a 99:1 split ratio, ensuring that

all clinical records belonging to a single patient are confined entirely to the same subset. Furthermore, due to computational resource constraints, we filter the generated samples by their context length, exclusively retaining those with up to 6k tokens for the trainset and 8k tokens for the testset. Following this filtering process, the final trainset and testset comprise 115,191 samples and 4,242 samples, respectively. The proportions of the event types are illustrated in Appendix D.

To comply with the data use agreement of MIMIC-IV, all data curation, model training, and test processes are executed locally using open-source models. No proprietary or closed-source models are utilized.

5.2 Models

ClinSim is developed based on the Qwen3 family of models (Yang et al., 2025). Specifically, we utilize the 4B parameter variant as the foundation model for training. We employ Qwen3-235B-A22B (Qwen3-235B for short) as the teacher model to synthesize high-quality Chain-of-Thought (CoT) reasoning traces. For efficiency considerations, the parsing process described in Section 3.2 is initially executed by the highly efficient Qwen3-30B-A3B, and any instances that fail to parse are sequentially escalated to Qwen3-32B and ultimately Qwen3-235B.

We compare ClinSim against the Qwen3 family (Qwen3-4B, Qwen3-8B, Qwen3-32B, and Qwen3-235B) and several additional baseline models. DeepSeek-R1-0528-Qwen3-8B distilled from DeepSeek-R1-0528 (Guo et al., 2025), GPT-OSS-20B, and GPT-OSS-120B (OpenAI et al., 2025) are included as strong general-purpose baselines. HuatuoGPT-o1 series (Chen et al., 2025) and Baichuan-M2-32B (Baichuan-M2 Team et al., 2025) are included as state-of-the-art medical QA models. General-purpose models and medical QA models are evaluated in a few-shot manner with one output example for each category in the prompt.

5.3 Data Synthesis and Training

To instill robust diagnostic reasoning capabilities, we construct an augmented trainset using synthetic CoT traces generated by a teacher model. Given the input context and the ground-truth answer, the teacher model first performs unrestricted internal analysis within `<think>` and `</think>` tags, where hindsight bias is naturally permitted, and then produces a distilled reasoning trace within

`<clean_think>` and `</clean_think>` tags. We instruct the model to justify the examination result based solely on preliminary information in the clean reasoning trace, and only extract this part to train our student model, ensuring the reasoning is completely free of hindsight bias and accurately reflects a clinician’s prospective decision-making. Finally, the ground-truth answer is appended to the clean reasoning trace to form the final training target for the student model.

Building upon this data, we design a two-stage training pipeline to develop the final ClinSim model. We divide the dataset into two patient-level subsets of equal size. In the first stage, we utilize the CoT-augmented samples from the first half of the dataset to perform SFT on Qwen3-4B. This cold-start phase efficiently equips the model with fundamental clinical knowledge and instills the distilled CoT reasoning paradigm. In the second stage, we initialize the model with the resulting cold-start checkpoint and apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) using the second half of the dataset. During GRPO, the model generates multiple candidate responses for each clinical context, and the policy is updated using a composite reward signal consisting of the response’s SimScore (0.9 weight) and a binary formatting reward for valid JSON (0.1 weight). The detailed SFT and GRPO hyperparameters are provided in Appendix A, while Appendix G provides the detailed instructions and prompt templates, including the CoT distillation and examination feedback simulation prompts relevant to this training pipeline.

5.4 Main Results

The experimental results are summarized in Table 1, with performance evaluated by SimScore. ClinSim surpasses all baseline models by at least 11.34 percentage points in overall SimScore and 20.97 percentage points in macro SimScore. Qwen3-32B outperforms other baselines in almost all metrics, except for *Microbiology* events, where it is surpassed by HuatuoGPT-o1-72B. While our 4B-parameter ClinSim achieves an exceptionally low parsing error rate of 0.5%, the two largest baseline models (Qwen3-235B and GPT-OSS-120B) achieve flawless formatting with a 0.0% ParseErr. This highlights the inherent advantage of massive parameter scale in maintaining strict structural consistency for complex outputs. Additionally, the HuatuoGPT-o1 series and Baichuan-M2 are built

Model	Radiology	Microbiology	Lab	Overall	Macro	ParseErr
<i>General Foundation Models</i>						
Qwen3-4B	0.2491	0.1586	0.2396	0.2336	0.2158	3.4%
Qwen3-8B	0.2603	0.1427	0.2623	0.2468	0.2218	0.2%
Qwen3-32B	<u>0.2667</u>	0.1752	<u>0.2905</u>	<u>0.2665</u>	<u>0.2441</u>	0.2%
Qwen3-235B	0.2619	0.1408	0.2827	0.2567	0.2285	0.0%
DeepSeek-R1-Qwen3-8B	0.2175	0.0872	0.2344	0.2093	0.1797	7.1%
GPT-OSS-20B	0.2634	0.1256	0.2538	0.2420	0.2143	0.2%
GPT-OSS-120B	0.2642	0.1604	0.2755	0.2567	0.2334	<u>0.0%</u>
<i>Medical Domain Models</i>						
HuatuoGPT-o1-7B	0.0780	0.1074	0.2094	0.1425	0.1316	29.7%
HuatuoGPT-o1-8B	0.0889	0.1091	0.1397	0.1149	0.1126	22.6%
HuatuoGPT-o1-70B	0.0298	0.1489	0.2557	0.1492	0.1448	37.2%
HuatuoGPT-o1-72B	0.2148	<u>0.2280</u>	0.2534	0.2344	0.2321	10.3%
Baichuan-M2-32B	0.2546	0.1297	0.2600	0.2418	0.2148	5.7%
<i>Ours</i>						
ClinSim-SFT	0.3283	0.4748	0.2548	0.3122	0.3526	12.9%
ClinSim-CoT	0.3300	0.7035	0.3016	0.3627	0.4451	3.3%
ClinSim-GRPO	0.3245	0.2380	0.2984	0.3018	0.2870	0.5%
ClinSim	0.3553	0.6850	0.3210	0.3799	0.4538	0.5%

Table 1: The performance evaluated using SimScore. The Overall and Macro scores represent the mean SimScore across all samples and across the three categories, respectively. The best scores are highlighted in **bold**, and the best scores among the baselines are underlined.

upon previous-generation foundation models such as Llama 3.1 (Grattafiori et al., 2024) and Qwen 2.5 (Qwen et al., 2025). Older foundations potentially limit their precision in strict structural generation, contributing to their notably higher ParseErr.

The results reveal varying difficulty across event types. Although absolute scores cannot be directly compared across types because each category has different components in SimScore, we can still analyze the task complexity with the relative performance gains over the baselines.

Lab events emerge as the most challenging domain to simulate. Compared to the best-performing baseline (Qwen3-32B), ClinSim achieves a modest relative improvement of approximately 10.5%. This difficulty stems from the vast combinatorial space of lab panels and the strictness of the Lab Scorer, which heavily penalizes omitted test items, fabricated labels, and numeric inaccuracies in physiological values. In contrast, *Microbiology* events demonstrate the highest learnability. While baseline models struggle severely (peaking at 0.2280), ClinSim achieves a significant increase to 0.6850. This suggests that *Microbiology* results, often characterized by standardized patterns (e.g., “no growth” or specific pathogen-antibiotic sensitivity tuples), can be fit by the model during training. *Radiology* events exhibit moderate difficulty, with Clin-

Sim demonstrating 33.2% improvement over the strongest baseline. Because radiology reports consist of free-text descriptions, the performance is intrinsically bounded by the natural variability in clinical phrasing.

Furthermore, consistent with the SimScore evaluation, ClinSim also achieves state-of-the-art performance across multiple traditional NLP metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and Levenshtein similarity (Levenshtein, 1965), with Qwen3-32B and Qwen3-235B remaining the strongest baselines. The detailed NLP performance is summarized in Appendix C.

5.5 Ablation Studies

Model Variant	Data: 1st Half	Data: 2nd Half
ClinSim-SFT	SFT-QA (Full Dataset)	
ClinSim-CoT	SFT-CoT (Full Dataset)	
ClinSim-GRPO	GRPO (Full Dataset)	
ClinSim	SFT-CoT	GRPO

Table 2: Ablation configurations for training strategies. The dataset is evenly divided into two patient-level subsets. The hybrid model (ClinSim) sequentially utilizes CoT distillation on the first half, followed by GRPO alignment on the second half.

As detailed in Table 2, we conduct an ablation

study evaluating three alternative training configurations. The results in Table 1 offer critical insights into the contribution of each training component.

The suboptimal performance of ClinSim-SFT demonstrates the necessity of reasoning. The introduction of synthetic reasoning traces (ClinSim-CoT) dramatically reduces the ParseErr to 3.3% and improves all domain scores. This confirms that CoT acts as a crucial cognitive bridge and format stabilizer, decomposing the complex predictive task into manageable logical steps.

The results of ClinSim-GRPO reveal the need for a cold-start phase. Although pure GRPO achieves higher scores than the strongest baselines on *Radiology* and *Lab* events, it struggles to improve the performance of *Microbiology* events (0.2380 in SimScore), showing virtually no advantage over the best baseline (0.2280 in SimScore). We attribute this to the specific pattern of *Microbiology* events, which causes reward sparsity. Without prior exposure to this distribution, the model can hardly discover high-reward generation trajectories.

The final ClinSim model, which harmonizes a CoT-based SFT cold-start with subsequent GRPO refinement, achieves better overall performance than ClinSim-CoT. It is hypothesized that even if the curated CoT samples are well-designed, they are still: (1) generated by a teacher LLM that is significantly larger than the target student model, introducing a capacity gap, and (2) synthesized retrospectively from the structured gold records, leading to a potential distributional shift. These factors may lead to reasoning patterns that are sub-optimal when generalized to the specific parameter space of the 4B model. The subsequent GRPO stage may partially mitigate this by allowing the model to autonomously adapt and polish its reasoning trajectories beyond the static limitations of the teacher’s synthetic data.

It is also observed that ClinSim-CoT, trained without SimScore-guided GRPO, already substantially outperforms the baselines, while the subsequent GRPO stage provides only a modest additional gain. This indicates that CoT distillation provides the main task adaptation, while GRPO provides a small final refinement in both output validity and average content score. Together with the improvements on traditional NLP metrics reported in Appendix C, these results suggest that the performance gains are not solely attributable to direct optimization of SimScore.

5.6 Validation of SimScore

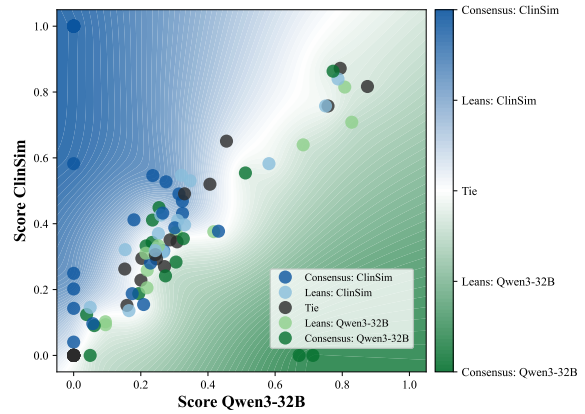


Figure 4: Continuous distribution of SimScore against human physician consensus. The background color gradient represents the intensity of human preference, while scatter markers indicate individual case assessments.

To preliminarily assess the alignment of SimScore with expert judgment, we conducted an independent, blinded human evaluation on 100 randomly sampled cases. Two medical experts were invited to evaluate the generated events from ClinSim against Qwen3-32B, the best-performing baseline.

Three choices were provided for each case: Model A is better, Model B is better, or Tie (indicating that both models performed equally well or equally poorly). The ground-truth examination report was provided as a reference for 50 cases, while it was intentionally masked for the remaining 50. To maintain blinding, Model A and Model B were randomly assigned to either ClinSim or Qwen3-32B for each case. Based on the combination of choices, the evaluations were categorized into an ordinal consensus gradient. As shown in Figure 4, **Consensus** refers to cases where both experts agree on a distinct preference, **Leans** indicates a weak consensus where one prefers a model while the other chooses Tie, and **Tie** denotes unanimous ties.

The experts exhibited general agreement, conflicting on only 11 samples (which were treated as **Tie** in Figure 4, but excluded from ordinal correlation analysis to maintain strict gradient validity). Overall, SimScore demonstrates a positive rank correlation (Spearman’s $r = 0.529$, $p < 0.001$ (Spearman, 1987)) and a reasonable ordinal concordance (C-index = 0.699 (Harrell et al., 1982)) with the intensity gradient of human clinical preferences. These results provide preliminary evidence of a

generally positive alignment between SimScore and physician judgment.

6 Conclusion

In this study, we formulate the task of examination feedback simulation to address the critical challenge of clinical trajectory uniqueness in developing realistic clinical simulation environments. We introduce the large-scale ClinTrack dataset and the structure-aware evaluation metric SimScore as the supporting tools. Experiments reveal that existing models struggle on the task. To bridge this gap and validate the feasibility of our formulation, we developed ClinSim as a proof-of-concept generative model. Results demonstrate that ClinSim significantly outperforms massive foundation models by over 10 percentage points in overall SimScore, indicating that the capability to simulate medically plausible, structured examination outcomes is highly trainable, laying the foundation for creating dynamic longitudinal diagnosis-and-treatment simulation frameworks for training and evaluating advanced clinical decision-making LLMs.

Limitations

We summarize the limitations of our work as follows:

- **Unexplored Counterfactual Clinical Trajectories:** While our evaluation provides a scalable, reference-anchored proxy for the intended off-path setting, direct evaluation on genuinely off-path and unrecorded examinations remains outside the scope of the current benchmark. Clinician-designed counterfactual scenarios are an important direction for future work.
- **Data and Modality Constraints:** ClinTrack relies on a single-center database (MIMIC-IV), which may reflect institution-specific ordering habits, documentation styles, and regional practice patterns. Future multi-institutional validation is needed to assess the broader generalizability of ClinTrack and ClinSim. Additionally, the current scope is strictly confined to textual and tabular inpatient EHR data, thereby excluding multi-modal clinical inputs such as radiological images or continuous ECG waveforms.

- **Validity of Metric:** Higher SimScore values do not by themselves guarantee broader clinical quality or downstream clinical utility. Further validation of SimScore is needed to ensure its reliability (e.g., finer-grained preference scale to avoid difficulties in capturing nuanced model differences, such as numerical precision versus textual coherence). Using SimScore as the reward as well as the evaluation metric may also introduce metric circularity.
- **Hyperparameter Optimization:** Due to computational constraints, we did not perform an exhaustive grid search for the two-stage training pipeline. While our results demonstrate the feasibility of this alignment strategy, further hyperparameter tuning and detailed ablation studies could lead to deeper insights into the training dynamics of this task.
- **Finer-grained Evaluation Metrics for Radiology:** Radiology reports are currently evaluated using purely text-based metrics. Future work should explore finer-grained rubric-based and structure-aware evaluation.

Ethical Considerations

Our ClinTrack dataset is curated from the MIMIC-IV database, which has been rigorously de-identified to protect patient privacy. To strictly adhere to the MIMIC-IV data use agreement, all data curation, model training, and evaluation processes were executed locally using open-source models. No proprietary or closed-source cloud APIs were utilized to avoid the risk of leakage. Finally, AI assistants were employed only for language polishing of the manuscript and did not influence the scientific content of this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576107); the Science and Technology Commission of Shanghai Municipality (No. 24511104200); Shanghai Municipal Commission of Economy and Informatization, “Corpus Construction for Large Language Models in Pediatric Respiratory Diseases” (No. 2024-GZL-RGZN-01013); Shanghai Municipal Health Commission, “Early Warning and Auxiliary Diagnosis and Treatment of Pediatric Severe

Pneumonia” (No. 2025ZHYL040); Municipal Hospital Frontier Joint Research Project, “Study on Evaluation Indicator Construction and Clinical Application Management for Assisted Diagnostic and Treatment Large Language Model in Pediatric Severe Pneumonia” (No. SHDC12024136); “Applying and Evaluating a Medical Large Language Model for the Diagnosis and Treatment of Pediatric Severe Pneumonia” (No. EKYX202409); and “Noncommunicable Chronic Diseases-National Science and Technology Major Project” (Grant No. 2025ZD0552303).

References

- Fatemeh Amrollahi, Nicholas Marshall, Fateme Nateghi Haredasht, Kameron C. Black, Aydin Zahedivash, Manoj V. Maddali, Stephen P. Ma, Amy Chang, Stanley C. Deresinski, Mary Kane Goldstein, Steven M. Asch, Niaz Banaei, and Jonathan H. Chen. 2024. A multi-phase analysis of blood culture stewardship: Machine learning prediction, expert recommendation assessment, and LLM automation. *AMIA Annual Symposium Proceedings*, 2024:147–156.
- Baichuan-M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. *Baichuan-m2: Scaling medical capability with large verifier system*. Preprint, arXiv:2509.02208.
- Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdounour, Adrian D. Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Liam G. McCoy, Haadi Mombini, Christopher Lucas, Misha Fotoohi, Matthew Gwiazdon, and 6 others. 2025. *Superhuman performance of a large language model on the reasoning tasks of a physician*. Preprint, arXiv:2412.10849.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. *Towards medical complex reasoning with LLMs through medical verifiable problems*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. *Ddxplus: A new dataset for automatic medical diagnosis*. In *Advances in Neural Information Processing Systems*, volume 35, pages 31306–31318. Curran Associates, Inc.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. *Deepseek-r1 incentivizes reasoning in llms through reinforcement learning*. *Nature*, 645(8081):633–638.
- Ronald M Harden, Mary Stevenson, W Wilson Downie, and GM Wilson. 1975. *Assessment of clinical competence using objective structured examination*. *Br Med J*, 1(5955):447–451.
- Frank E. Harrell, Jr., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. *Evaluating the yield of medical tests*. *JAMA*, 247(18):2543–2546.
- Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. 2023. *MIMIC-IV, a freely accessible electronic health record dataset*. *Scientific Data*, 10(1):1.
- Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. 2019. *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports*. *Scientific Data*, 6(1):317.
- Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. *CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs*. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- William L. Kissick. 1994. *Medicine’s Dilemmas: Infinite Needs versus Finite Resources*. Yale University Press.
- Vladimir I. Levenshtein. 1965. *Binary codes capable of correcting deletions, insertions, and reversals*. *Soviet physics. Doklady*, 10:707–710.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. *Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning*. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025a. [Interactive evaluation for medical LLMs via task-oriented dialogue system](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju, Shizheng Li, and Jian Xie. 2025b. [Exploring the inquiry-diagnosis relationship with advanced patient simulators](#). *Preprint*, arXiv:2501.09484.
- Luyang Luo, Sung Eun Kim, Xiaoman Zhang, Julius M Kernbach, Roshan Kenia, Julian N Acosta, Larry A Nathanson, Adrian D Haimovich, Adam Rodman, Ethan Goh, Jonathan H Chen, Nigam H Shah, David A Kim, James Zou, Faisal Mahmood, Jakob Nikolas Kather, Matthew Lungren, Vivek Natarajan, Eric J Topol, and Pranav Rajpurkar. 2026. [A clinical environment simulator for dynamic AI evaluation](#). *Nature Medicine*, 32(3):820–827.
- Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. 2025. [Sequential diagnosis with language models](#). *Preprint*, arXiv:2506.22405.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [Agentclinic: a multimodal benchmark for tool-using clinical ai agents](#). *npj Digital Medicine*.
- Tianqi Shang, Weiqing He, Charles Zheng, Lingyao Li, Li Shen, and Bingxin Zhao. 2025. [Dynamicare: A dynamic multi-agent framework for interactive and open-ended medical decision-making](#). *Preprint*, arXiv:2507.02616.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.
- C. Spearman. 1987. [The proof and measurement of association between two things](#). *The American Journal of Psychology*, 100(3/4):441–471.
- Zhoujian Sun, Ziyi Liu, Cheng Luo, Jiebin Chu, and Zhengxing Huang. 2025. [Improving interactive diagnostic ability of a large language model agent through clinical experience learning](#). *Preprint*, arXiv:2503.16463.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. [TRL: Transformer Reinforcement Learning](#). <https://github.com/huggingface/trl>.
- Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. [CMB: A comprehensive medical benchmark in Chinese](#). In *Proceedings of the*

2024 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Ahmed Medhat Zayed, Glynis Frans, and Nicolas Delvaux. 2025. [Evaluating large language models as clinical laboratory test recommenders in primary and emergency care: a crucial step in clinical decision making](#). *Clinical Chemistry and Laboratory Medicine (CCLM)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Yuxuan Zhou, Xien Liu, Chenwei Yan, Chen Ning, Xiao Zhang, Boxun Li, Xiangling Fu, Shijin Wang, Guoping Hu, Yu Wang, and Ji Wu. 2025. [Evaluating LLMs across multi-cognitive levels: From medical knowledge mastery to scenario-based problem solving](#). In *Forty-second International Conference on Machine Learning*.

A Experimental Setup Details

During training, all models are initialized from the Qwen3-4B base model. The training process is conducted using the TRL (von Werra et al., 2020) library on eight GPUs, with DeepSpeed optimization employed to manage memory constraints. For the hybrid two-stage training strategy (ClinSim) detailed in Section 5.5, the dataset subsets (halves) are each trained for a full epoch sequentially, rather than splitting a single epoch across them. Due to computational constraints, the reported model performances are derived from a single experimental run. Greedy decoding is used for all inference generations at test time.

The detailed hyperparameter configurations for both the SFT and GRPO stages are summarized in Table 3.

Hyperparameter	QA & CoT	GRPO
Optimizer	AdamW	AdamW
Learning rate	1e-5	1e-6
Scheduler	Cosine	Cosine
Warmup ratio	0.05	0.01
Precision	bf16	bf16
Max length	8192	8192
Per-device batch size	1	1
Gradient accumulation steps	64	64
Generations per prompt	-	4
Temperature	-	1.0
Format reward weight	-	0.1
SimScore reward weight	-	0.9
KL coefficient (β)	-	1e-3
Clip range (ϵ)	-	0.2

Table 3: Hyperparameter configurations for different training stages.

B Human Audit of the Parsing Pipeline

To quantify the potential label noise introduced by our LLM-based parsing pipeline, we conducted a comprehensive human audit covering both the discharge note extraction and radiology report parsing tasks.

B.1 Audit of Discharge Note Extraction

For the structured extraction of discharge notes, we randomly sampled 100 cases from our dataset and manually annotated the ground-truth JSONs. The LLM-generated outputs were then compared against these human-annotated references. We measured the agreement using BLEU, ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), and Levenshtein similarity. The results, summarized in Table 4, show high agreement between the automated

extraction and the human annotations. These scores confirm that our parsing pipeline is highly accurate, ensuring that the structured information faithfully represents the original clinical notes.

Metric	Score
BLEU	85.24%
ROUGE-1	93.08%
ROUGE-2	88.26%
ROUGE-L	92.26%
Levenshtein Similarity	95.78%

Table 4: Average agreement metrics between the LLM-generated structured extractions and human expert annotations on 100 randomly sampled discharge notes.

B.2 Audit of Radiology Report Parsing

For the radiology report parsing task, we performed a stratified sampling of 120 reports to ensure balanced coverage. Specifically, we randomly selected 20 samples for each of the six primary modalities: *X-ray*, *CT*, *Ultrasound*, *MRI*, *Fluoroscopy*, and *Mammography*.

Given that radiology reports typically contain explicit modality and body part keywords in their textual descriptions, this parsing task is relatively straightforward. The human audit confirmed this observation, yielding an accuracy of 100% across all 120 sampled cases. This perfect alignment indicates that the labels used in our dataset are highly reliable, with a negligible risk of label noise affecting downstream training or evaluation.

C Traditional NLP Metrics

In addition to the SimScore evaluation, we also assess model performance using traditional NLP metrics, including BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), Levenshtein similarity (Levenshtein, 1965), and BERTScore (Zhang et al., 2020) computed with RoBERTa-Large (Liu et al., 2019). The detailed results are presented in Table 5.

D Event Type Proportion

Figure 5 shows the detailed distribution of event types and their respective subtypes within the trainset and testset. To offer a deeper insight into the data diversity, the subsequent figures (Figure 6, Figure 7, and Figure 8) provide a fine-grained breakdown of modalities and clinical categories specifically within the trainset.

Model	BLEU	R-1	R-2	R-L	Lev	BERTScore
<i>General Foundation Models</i>						
Qwen3-4B	0.5401	0.5109	0.3544	0.4412	0.5635	0.9115
Qwen3-8B	0.5327	0.5188	0.3568	0.4404	0.5705	0.9136
Qwen3-32B	0.5596	<u>0.5415</u>	0.3799	<u>0.4650</u>	<u>0.5893</u>	<u>0.9181</u>
Qwen3-235B	<u>0.5701</u>	<u>0.5400</u>	<u>0.3820</u>	<u>0.4602</u>	<u>0.5887</u>	<u>0.9175</u>
DeepSeek-R1-Qwen3-8B	0.5410	0.5029	0.3494	0.4275	0.5489	0.9047
GPT-OSS-20B	0.5282	0.5256	0.3550	0.4439	0.5768	0.9124
GPT-OSS-120B	0.5359	0.5375	0.3717	0.4522	0.5814	0.9138
<i>Medical Domain Models</i>						
HuatuoGPT-o1-7B	0.5132	0.3820	0.2841	0.3482	0.4243	0.8762
HuatuoGPT-o1-8B	0.4018	0.3956	0.2857	0.3579	0.4508	0.8812
HuatuoGPT-o1-70B	0.4256	0.3651	0.2894	0.3378	0.3969	0.8695
HuatuoGPT-o1-72B	0.5428	0.5025	0.3700	0.4488	0.5518	0.9087
Baichuan-M2-32B	0.5423	0.5037	0.3528	0.4315	0.5501	0.9052
<i>Ours</i>						
ClinSim-SFT	0.3780	0.5059	0.3941	0.4624	0.5473	0.9064
ClinSim-CoT	0.5038	0.5863	0.4644	0.5367	0.6275	0.9287
ClinSim-GRPO	0.5657	0.5797	0.4192	0.5086	0.6230	0.9252
ClinSim	0.5786	0.6137	0.4690	0.5493	0.6560	0.9350

Table 5: The performance evaluated using traditional NLP metrics. The best scores are highlighted in **bold**, and the best scores among the baselines are underlined. R-1, R-2, and R-L denote the ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. Lev denotes the Levenshtein similarity. BERTScore is computed with RoBERTa-Large.

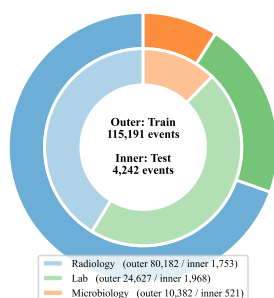


Figure 5: Overall distribution of events in the trainset and testset.

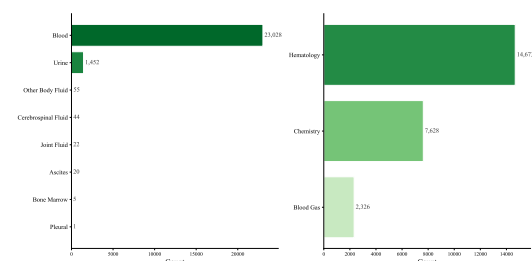


Figure 7: Distribution of laboratory event subtypes by fluid and category within the trainset.

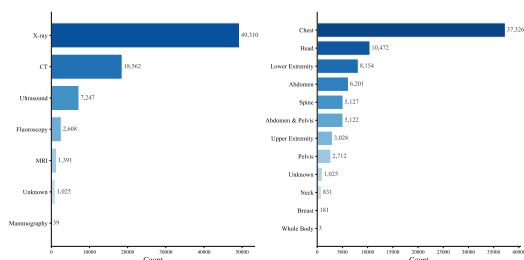


Figure 6: Distribution of radiology event subtypes by modality and body part within the trainset.

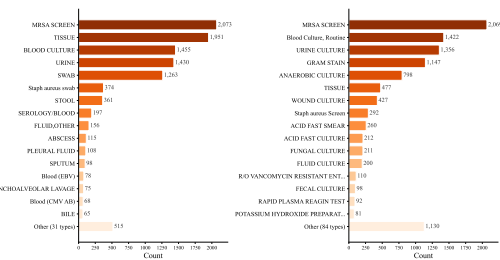


Figure 8: Distribution of microbiology event subtypes by specimen type and test name within the trainset.

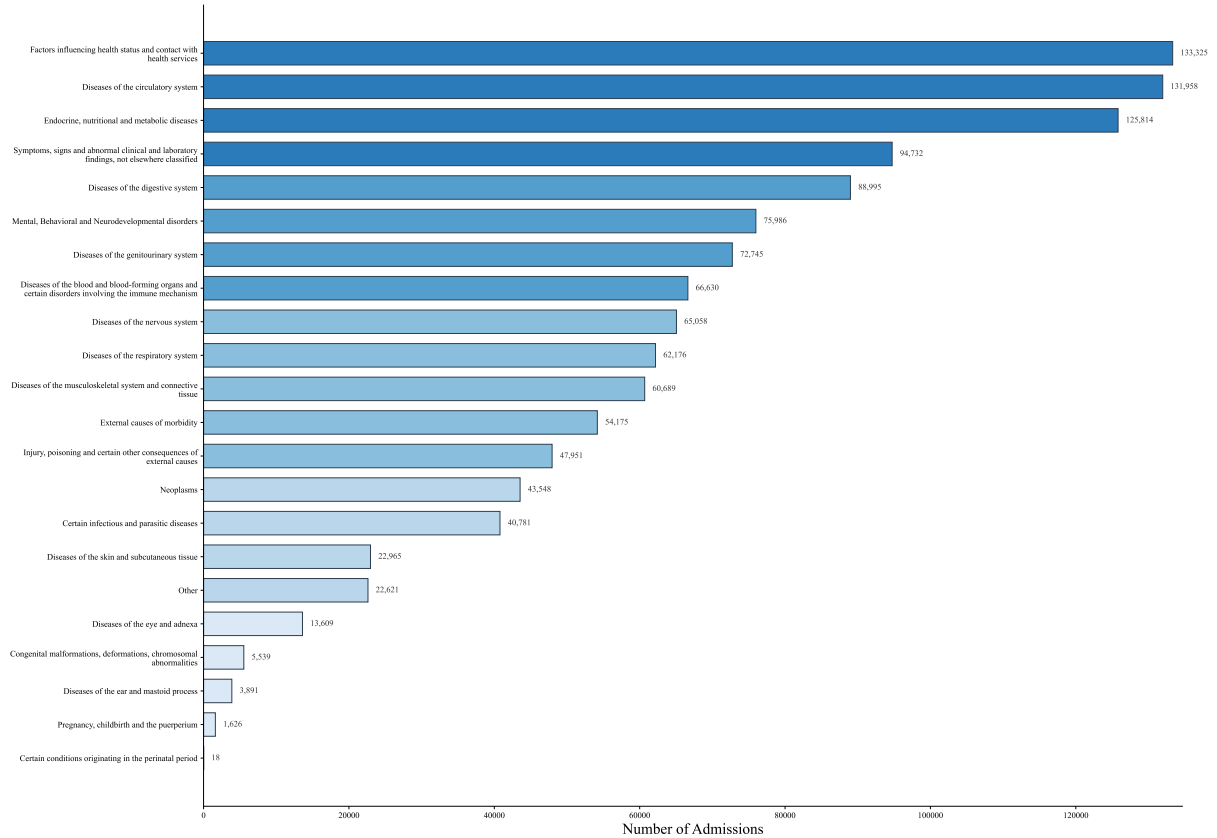


Figure 9: ICD-10 chapter-level disease distribution across all admissions.

E Disease Distribution

We analyzed the disease distribution across ~188K admissions from ~103K patients using ICD-10 chapter-level classification. For each admission, we counted each chapter at most once regardless of how many diagnosis codes fell within it, so the percentages reflect the proportion of admissions in which a given disease category appeared. Figure 9 shows the full chapter-level distribution.

F Effect of Final-Diagnosis Conditioning

To examine whether ClinSim relies on the diagnosis field as a shortcut, we conducted an inference-time ablation using the same ClinSim checkpoint and test set. The results are presented in Table 6.

These results suggest that ClinSim’s performance is not primarily driven by directly exploiting the final diagnosis field. The longitudinal ClinTrack input already contains rich disease-state cues through past medical history and preceding clinical events, which may make the final diagnosis partly redundant under this MIMIC-IV distribution.

Nevertheless, when modeling, we retain the final diagnosis list as a clinically grounded auxiliary

anchor for specifying the patient state, because the degree of redundancy between longitudinal history and final diagnoses may vary across hospital datasets.

G Prompt Templates

We present the prompt templates utilized across different stages of our framework from page 17 to page 18, containing instructions for Chain-of-Thought distillation, examination feedback simulation, as well as discharge and radiology note parsing. To preserve privacy and comply with data usage agreements, patient-specific information in the few-shot examples has been replaced with placeholders.

H Data Samples

MIMIC-IV and its derivatives have strict data usage policies that prevent us from sharing actual patient data. Therefore, we created a brief fictitious example that mimics the structure and content of real clinical data while ensuring patient privacy (page 19).

Diagnosis condition	Macro	Overall	Lab	Radiology	Microbiology	ParseErr
Included	0.4538	0.3799	0.3210	0.3553	0.6850	0.5%
Removed	0.4530	0.3804	0.3249	0.3535	0.6807	0.4%

Table 6: The effect of final-diagnosis conditioning on SimScore performance.

I Details of SimScore Validation

Two medical experts with clinical experience were invited to participate in the human validation of SimScore. Their participation was managed through academic collaboration and governed by internal research standards. The full instructions provided to the experts are detailed at the end (page [21](#)).

CoT Distillation Instructions

You are a medical data annotation expert. Your task is to write high-quality "reasoning chain-of-thought" training data for a smaller AI model.

Your Workflow

1. You will receive a medical prediction task consisting of a [Question] and its corresponding [Reference Answer].
2. Carefully read the patient information and clinical context provided in the question.
3. Use the reference answer to understand the correct reasoning direction.
4. Then, write a **clean reasoning process** that:
 - Starts from the question and systematically analyzes the patient's condition
 - Applies medical knowledge for logical deduction
 - Naturally arrives at a conclusion consistent with the reference answer
 - **MUST NEVER** mention "reference answer", "given answer", "ground truth", "we know the answer is", or any phrasing that implies prior knowledge of the result
5. **Critical Anti-Leakage Rule**: Your reasoning must **ONLY** use information available in the [Question]. You must **NOT** incorporate any specific details from the [Reference Answer] into your reasoning. This includes but is not limited to:
 - Specific test results, lab values, or imaging findings from the answer
 - Exact wording or phrases that appear in the answer's report text
 - Specific diagnoses or conclusions that are only revealed in the answer
 - Any numeric values (e.g., counts, measurements) from the answerThe reference answer is provided solely so you understand what the task is asking for and what direction the reasoning should take.
6. **Temporal Boundary Rule**: Your reasoning must end **BEFORE** the examination is performed. You must **NOT** describe, summarize, or comment on the outcome or results of the examination in any way - not even in general terms such as "the findings were normal", "no abnormalities were found", or "the results confirmed the diagnosis". The reasoning should explain **WHY** an exam is ordered and **WHAT** the clinician expects to find, but must never state what was actually found. Think of yourself as the clinician writing a pre-exam clinical note.

Output Format

Wrap your clean reasoning process with `<clean_think>` and `</clean_think>` tags.
Do NOT output anything outside the tags.

Example of CORRECT reasoning (no leakage)

`<clean_think>`

The patient is a 27-year-old woman with Crohn's disease presenting with acute abdominal pain, nausea, and vomiting, with a history of prior SBO episodes and ileocecectomy. The acute worsening of symptoms and pain progression raise concern for recurrent small bowel obstruction. An abdominal X-ray is the standard initial imaging modality in suspected SBO, as it can quickly assess for bowel distention patterns, air-fluid levels, and free air. Given the clinical urgency and her surgical history, this is the logical first-line diagnostic step before considering advanced imaging such as CT.

`</clean_think>`

Examples of WRONG reasoning (DO NOT do this)

WRONG Example 1 - Leaking specific findings from the results:

"The radiographic findings of scattered air-fluid levels and gaseous distention in the mid-to-lower abdomen support partial SBO."

Describes specific imaging findings that only exist in the exam results.

WRONG Example 2 - Summarizing the outcome, even in general terms:

"The imaging findings showed no acute intrathoracic abnormalities, which helped exclude life-threatening respiratory causes."

States what the exam showed. The reasoning should stop BEFORE the results are known.

Examination Feedback Simulation Instructions

You are a medical prediction assistant tasked with forecasting the potential results of a patient's subsequent medical examination based on their existing data.

Instructions

1. **Predict**: You should predict the full details of the event.
2. **Format**: Your final response must be a valid JSON object strictly following the format of the examples below. Do not include any text outside the JSON block.

Input Format

1. Preliminary Information: A chronological list of past events.
2. Auxiliary Diagnosis List: The patient's final diagnosis list, provided only as simulator-side case information.
3. Information to Predict: The metadata of the event to predict.

Reference Examples

The following examples demonstrate the required JSON structure for different event types. Note that these examples only show the final JSON output.

{examples_block}

{examples_block} denotes a runtime-injected block containing one output-format example for each event type. The exact block is provided in the released repository.

Discharge Note Parsing Instructions

You are a medical assistant. You are responsible for parsing the following discharge note and extracting the relevant information in the json format.

You should strictly copy the information from the discharge note and paste it to corresponding fields.

If any information is missing, please leave the field empty by filling it with null.

The json format should be as follows:

```
{
  "Admission Info": {
    "Allergies": "text",
    "Chief Complaint": "text",
    "History of Present Illness": "text",
    "Past Medical History": "text",
    "Social History": "text",
    "Family History": "text",
    "Admission Physical Exam": "text",
    "Medications on Admission": "text"
  },
  "Progress Info": {
    "Major Surgical or Invasive Procedure": "text",
    "Pertinent Results": "text",
    "Brief Hospital Course": "text"
  },
  "Discharge Info": {
    "Discharge Condition": "text",
    "Discharge Physical Exam": "text",
    "Discharge Instructions": "text",
    "Discharge Diagnosis": "text",
    "Discharge Medications": "text",
    "Followup Instructions": "text"
  }
}
```

Note: all "text" should be replaced with the actual text from the discharge note in plain text format. The answer should not include any type of other elements like "```json".

Radiology Note Parsing Instructions

You are a medical assistant specialized in Radiology.

Your task is to classify the **Modality** and **Body Part** of a radiology note based on the following standards.

All classification results must be chosen from the predefined valid labels listed below.

Target Labels

Valid Modalities: ['CT', 'MRI', 'Ultrasound', 'X-ray', 'Mammography', 'Fluoroscopy', 'Nuclear Medicine', 'Others']

Valid Body Parts: ['Head', 'Neck', 'Chest', 'Breast', 'Abdomen', 'Pelvis', 'Abdomen & Pelvis', 'Spine', 'Upper Extremity', 'Lower Extremity', 'Whole Body', 'Others']

Classification Rules

1. Modality Rules

- **CT**: Includes its angiographic variants (CTA) and venography (CTV).
- **MRI**: Includes MRA and MRV.
- **Ultrasound**: Includes "Sonogram", "Doppler", "Duplex", and "Echocardiogram".
- **X-ray**: Includes all plain radiographs (e.g., "Chest X-ray", "KUB", "Extremity X-ray").
- **Mammography**: Includes screening and diagnostic mammograms, as well as breast tomosynthesis.
- **Fluoroscopy**: Includes contrast studies (e.g., "Barium Swallow"), real-time imaging, Angiography (DSA/Angio), and C-Arm guidance for procedures (e.g., line placement).
- **Nuclear Medicine**: Includes all scans using radiotracers (e.g., "Bone Scan", "PET-CT", "HIDA", "V/Q Scan").

2. Body Part Rules

- **Head**: Includes Brain, Skull, Face, Sinus, Orbits.
- **Neck**: Soft tissue, Thyroid, Carotids (Exclude Spine).
- **Spine**: Any exam of the vertebrae (Cervical, Thoracic, Lumbar).
- **Spine Priority**: Any exam of the vertebrae (Cervical, Thoracic, Lumbar) is "Spine".
 - **Contrast**: "CT Neck" (Soft tissue) -> **Neck**. "CT C-Spine" (Bone) -> **Spine**.
- **Abdomen & Pelvis**:
 - Select this **ONLY** if the note explicitly covers **BOTH** regions (e.g., "CT Abd/Pelv").
 - If it is "CT Abdomen" only, select **Abdomen**.
 - If it is "CT Pelvis" only, select **Pelvis**.
- **Chest vs. Breast**:
 - "Chest" includes lungs, heart, mediastinum.
 - "Breast" is strictly for breast tissue imaging (Mammogram, Breast US/MRI).
- **Extremities (Upper vs. Lower)**:
 - **Upper Extremity**: Shoulder, Clavicle, Scapula, Humerus, Elbow, Forearm, Wrist, Hand, Finger.
 - **Lower Extremity**: Hip, Femur, Knee, Tibia, Fibula, Ankle, Foot, Toe.
 - **Note**: If a scan covers both (e.g., skeletal survey), select "Whole Body".

3. Conditions for "Others" Label

- **Not Listed**: The modality or body part is not in the provided valid labels.
- **Insufficient Information**: The text does not provide enough details to determine modality or body part.

- ****Multiple Conflicting Info****: The note mentions multiple modalities or body parts without a single, current focus. However, if one modality/body part is the current focus and the others are historical, classify as that primary one.

4. Standard Mapping

Treat "Computed Tomography" as "CT" and "Magnetic Resonance Imaging" as "MRI".

5. Current vs. Historical

You must distinguish the current examination from historical comparisons.

If a modality or body part is mentioned only in the context of prior studies, do NOT classify based on that.

Classify based solely on the examination currently being reported.

6. Output Format

You must output a JSON object with exactly two fields: "Modality" and "Body Part".

The output JSON should look like this:

```
{
  "Modality": "CT",
  "Body Part": "Chest"
}
```

Ensure the values for each field are selected from the predefined valid labels.

Example Data Entry

```
{
  "subject_id": 999999,
  "gender": "F",
  "birth": "2100",
  "dod": null,
  "admission": [
    {
      "hadm_id": 9000001,
      "admittime": "2100-01-01 10:00:00",
      "dischtime": "2100-01-04 12:00:00",
      "admission_type": "TEST OBSERVATION",
      "admit_provider_id": "P99TEST",
      "admission_location": "EMERGENCY ROOM",
      "insurance": "Private",
      "language": "English",
      "marital_status": "SINGLE",
      "race": "WHITE",
      "edregtime": "2100-01-01 08:30:00",
      "edouttime": "2100-01-01 11:00:00",
      "hospital_expire_flag": 0,
      "discharge_note": "This is a test discharge note for unit testing purposes. Content is non-clinical.",
      "diagnosis_list": [
        {
          "seq_num": 1,
          "icd_code": "TEST01",
          "icd_version": 99,
          "long_title": "Test Diagnosis A"
        },
        {
          "seq_num": 2,
          "icd_code": "TEST02",
          "icd_version": 99,
          "long_title": "Test Diagnosis B"
        }
      ]
    },
    {
      "admission_info": {
        "allergies": "Test Allergy (e.g., Testicillin)",
        "chief_complaint": "Test Complaint (e.g., Test Pain)",
        "history_of_present_illness": "Test HPI: Patient presents with test symptoms.",
        "past_medical_history": "Test PMH (e.g., Test Hypertension)",
        "social_history": null,
        "family_history": "Test Family History",
        "admission_physical_exam": "Test Physical Exam findings.",
        "medications_on_admission": "TestMed (e.g., Testopril)"
      },
      "progress_info": {
        "major_surgical_or_invasive_procedure": "Test Procedure (e.g., Test CABG)",
        "pertinent_results": "Test results are pending.",
        "brief_hospital_course": "Test hospital course. Patient stable."
      },
      "discharge_info": {
        "discharge_condition": "Test Condition (Stable)",
        "discharge_physical_exam": null,
        "discharge_instructions": "Test Discharge Instructions (e.g., Follow up)",
        "discharge_diagnosis": "Test Discharge Diagnosis (e.g., Test CAD)"
      }
    }
  ]
}
```

```

    "discharge_medications": "Test Discharge Meds (e.g., TestAspirin)",
    "followup_instructions": null
  },
  "event_list": [
    {
      "event_time": "2100-01-01T10:00:00",
      "event_type": "AdmissionEvent",
      "admission_type": "TEST_EMERGENCY",
      "admission_location": "TEST_LOCATION"
    },
    {
      "event_time": "2100-01-01T11:00:00",
      "event_type": "TransferEvent",
      "transfer_location": "Test Emergency Department"
    },
    {
      "event_time": "2100-01-01T12:00:00",
      "event_type": "ServiceEvent",
      "prev_service": null,
      "curr_service": "TEST_SURG"
    },
    {
      "event_time": "2100-01-02T10:00:00",
      "event_type": "LabEvent",
      "specimen_id": 99999999,
      "fluid": "Test Fluid (Blood)",
      "category": "Test Chemistry",
      "lab_list": [
        {
          "lab_event_id": 90001,
          "itemid": 50983,
          "value": "140",
          "valuenum": 140.0,
          "valueuom": "mEq/L",
          "ref_range_lower": 133.0,
          "ref_range_upper": 145.0,
          "flag": null,
          "priority": "STAT",
          "comments": null,
          "label": "Test Sodium"
        }
      ]
    },
    {
      "event_time": "2100-01-02T11:00:00",
      "event_type": "MicrobiologyEvent",
      "spec_type_desc": "Test Specimen (e.g., URINE)",
      "test_name": "Test Culture",
      "microbiology_list": [
        {
          "micro_event_id": 90002,
          "test_seq": 1,
          "org_itemid": null,
          "org_name": null,
          "isolate_num": null,
          "quantity": null,
          "ab_itemid": null,
          "ab_name": null,
          "dilution_text": null,
          "dilution_comparison": null,
          "dilution_value": null,
          "interpretation": null,
          "comments": "Test comments for microbiology."
        }
      ]
    },
    {
      "event_time": "2100-01-02T12:00:00",
      "event_type": "RadiologyEvent",
      "note_id": "TEST-NOTE-12345",
      "note_type": "RAD",
      "note_seq": 1,
      "text": "Test radiology report text. Findings are non-clinical.",
      "modality": "Test modality",
      "body_part": "Test body part"
    },
    {
      "event_time": "2100-01-02T13:00:00",
      "event_type": "PrescriptionEvent",
      "prescription_list": [
        {
          "starttime": "2100-01-02T13:00:00",
          "stoptime": "2100-01-03T13:00:00",
          "drug_type": "MAIN",

```

```

        "drug": "TestMed-A (e.g., Aspirin)",
        "formulary_drug_cd": "TESTDRUG",
        "gsn": null,
        "ndc": null,
        "prod_strength": "100mg Tablet",
        "form_rx": null,
        "dose_val_rx": "100",
        "dose_unit_rx": "mg",
        "form_val_disp": "1",
        "form_unit_disp": "TAB",
        "doses_per_24_hrs": 1.0,
        "route": "PO"
    }
]
},
{
    "event_time": "2100-01-02T14:00:00",
    "event_type": "EmarEvent",
    "emar_list": [
        {
            "emar_id": "999999-1",
            "emar_seq": 1,
            "medication": "TestMed-A (e.g., Aspirin)",
            "event_txt": "Administered"
        }
    ]
},
{
    "event_time": "2100-01-03T10:00:00",
    "event_type": "ProcedureEvent",
    "seq_num": 1,
    "icd_code": "TESTP1",
    "icd_version": 99,
    "long_title": "Test Procedure"
},
{
    "event_time": "2100-01-03T11:00:00",
    "event_type": "OmrEvent",
    "seq_num": 1,
    "result_name": "Test OMR",
    "result_value": "Test Value"
},
{
    "event_time": "2100-01-04T12:00:00",
    "event_type": "DischargeEvent",
    "discharge_location": "TEST HOME"
}
]
}
}
}
}

```

SimScore Human Validation Instructions

Task Overview

Welcome, and thank you for lending your medical expertise to this evaluation! Your task is to assess AI-simulated examination feedback to help us validate an automated scoring metric.

In each of the 100 cases, you will be provided with:

1. A patient's **prior clinical records** and **diagnosis list**.
2. A specified **examination type**.
3. Two simulated examination feedback samples generated by **Model A** and **Model B**.
4. **(Optional)** The **real report** for that visit as a reference (for 50 cases).

Evaluation Rules

- * **Blind A/B Comparison:** Based on your professional judgment, please choose the **better** report between Model A and Model B. Model identities are blinded and randomized across cases.
- * **Prioritize Clinical Plausibility:** The real report is provided merely for reference, not as a strict gold standard. Please evaluate the generated reports based on **medical reasonableness**. If an AI's report differs from the real outcome but remains clinically plausible and logical given the patient's context, it should still be considered highly acceptable.
- * **Edge Cases:** If you cannot confidently choose one over the other, please select the "Tie (both good/both bad)" option.