CRITIC–ADVISER–REVISER CYCLIC REFINEMENT: TOWARDS HIGH-QUALITY EMR CORPUS GENERATION WITH LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic medical records (EMRs) are vital for healthcare research, but their use is limited by privacy concerns. Synthetic EMR generation offers a promising alternative, yet most existing methods merely imitate real records without adhering to rigorous clinical quality principles. To address this, we introduce LLM-CARe, a stage-wise cyclic refinement framework that progressively improves EMR quality through three stages, each targeting a specific granularity: corpus, section and document. At each stage, a Critic, an Adviser, and a Reviser collaborate iteratively to evaluate, provide feedback, and refine the drafts. This structured, multi-stage process produces records that better satisfy clinical quality standards. Experiments show that LLM-CARe significantly enhances EMR quality across all levels compared to strong baselines and yields improved performance on real-world clinical tasks such as diagnosis prediction. Unlike prior work, our method requires no real EMRs for training or prompting, demonstrating the effectiveness of stage-wise, cyclic refinement for generating high-quality, privacy-preserving EMR datasets.

1 INTRODUCTION

Electronic Medical Records (EMRs) are a valuable resource for healthcare research (Ma et al., 2017; Shang et al., 2019; Shen et al., 2025), offering large-scale, clinically grounded insights that reflect real-world medical practice. However, the sensitive nature of patient information poses significant privacy challenges, which severely limit the open sharing and use of real EMRs (Iyengar et al., 2018; Shah & Khan, 2020; Tertulino et al., 2024). To mitigate these concerns, researchers have explored synthetic EMR generation methods that aim to preserve data utility while protecting patient confidentiality (Yan et al., 2022; Murtaza et al., 2023; Yuan et al., 2024).

Existing EMR synthesis approaches primarily focus on mimicking real records (Lee, 2018; Baowaly et al., 2018; Yoon et al., 2023), without explicit ensuring clinical soundness. This imitation-based strategy is risky: real EMRs may contain errors (Aerts et al., 2021; Mohamed et al., 2023), which can be inadvertently inherited by synthetic data (Figure 1(a)). In practice, EMRs are professional medical documents whose reliability depends on satisfying key requirements such as completeness, consistency, and distribution alignment. Synthetic records that fail to meet such requirements may be less useful—or even misleading—for downstream clinical or research applications.

Recent advances in large language models (LLMs) make them a promising tool for EMR generation, due to their text generation ability and rich internal knowledge. However, as shown in Figure 1(a), our preliminary analysis reveals that LLM outputs often exhibit biased distributions—such as unrealistic gender patterns—and insufficient coverage of less typical clinical cases. These challenges highlight the need for more structured approaches to harness LLMs effectively for EMR synthesis.

To bridge this gap, we propose **LLM**-based **C**ritic–**A**dviser–**R**eviser Cyclic R**e**finement (**LLM-CARe**), a stage-wise framework that enhances synthetic EMR quality through progressive refinement across corpus, section and document levels. As illustrated in Figure 1(b), LLM-CARe incorporates clinical quality principles into the generation process, producing records that align closely with professional standards of medical documentation. These requirements are organized into concrete principles of corpus distributional alignment, section completeness, and document consistency, forming the basis for refinement at different granularities. Guided by them, LLM-CARe proceeds

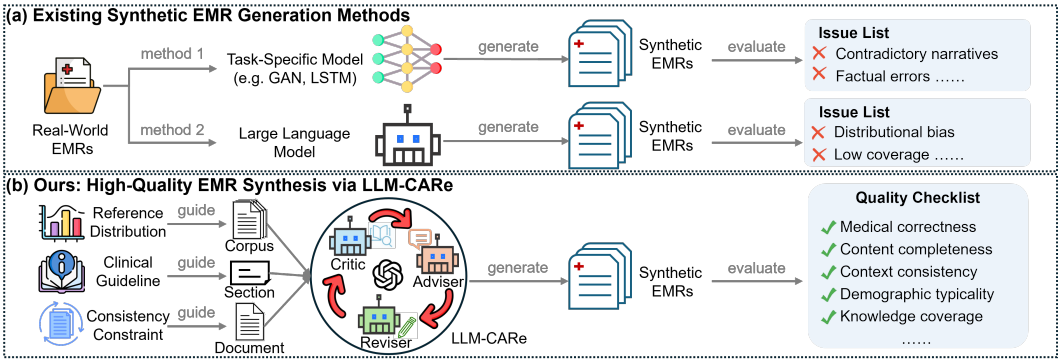

Figure 1: **(a)** Traditional EMR generation that mimic real EMRs without considering quality often leads to suboptimal outputs. **(b)** Our proposed LLM-CARe incorporates cyclic refinement based on quality principles, synthesizing high-quality EMRs.

through three stages of refinement—corpus, section, and document—each targeting a distinct aspect of EMR quality. Within every stage, a Critic, an Adviser, and a Reviser collaborate in a cyclic loop: the Critic evaluates the drafts, the Adviser provides targeted feedback, and the Reviser updates the records. While the interaction pattern is shared, the role of each agent adapts to the stage: corpus stage aligns the dataset with realistic distributions, section stage enforces section completeness, and document stage ensures logical consistency within record. This structured process enables systematic enhancement of EMRs from local detail to global corpus characteristics.

To validate the effectiveness of our approach, we conduct two types of evaluations on a real-world EMR dataset containing 192k records across 302 disease categories. We assess the intrinsic quality of generated records using a strong LLM as a judge, complemented by statistical comparisons to real EMRs. Additionally, we evaluate downstream utility by training task-specific models on synthetic EMRs and testing them on real records across three representative clinical tasks: diagnosis prediction, examination recommendation, and treatment recommendation. Results show that LLM-CARe consistently outperforms baseline methods in both record quality and task performance. Notably, our method requires no access to real EMRs during generation, fully preserving patient privacy while producing clinically meaningful and practically useful data.

Our main contributions are summarized as follows:

- We propose **LLM-CARe**, a stage-wise multi-agent framework for high-quality EMR synthesis that employs cyclic refinement based on clinical quality principles.
- LLM-CARe consistently improves EMR quality compared to baselines across multiple levels. Further analysis shows that all three agents play essential and complementary roles.
- Without using real EMRs, our synthetic data yields superior downstream task performance compared to baselines, ensuring both utility and privacy.

2 Related Work

There has been growing interest in synthesizing EMRs to address privacy concerns and facilitate secure data sharing. We categorize existing methods into three main paradigms:

GAN-based EMR Generation. Generative adversarial networks (GANs) have been extensively explored for EMR synthesis. Some works generate EMRs from random noise vectors (Choi et al., 2017; Baowaly et al., 2018; Chin-Cheong et al., 2019; Yoon et al., 2023), while other methods incorporate structured conditions—such as diagnosis codes—to guide generation process (Rashidian et al., 2020; Zhang et al., 2020; Guan et al., 2021; Li et al., 2023). Although effective at modeling data distributions, these methods typically ignore the clinical quality of the generated records.

Auto-regressive EMR Generation. Another line of research leverages auto-regressive models to generate EMRs. Recurrent neural networks (RNNs) have been used to model sequential EMR data

(Lee, 2018; Melamud & Shivade, 2019; Mosquera et al., 2023; Ganguli et al., 2023), and more recently, transformer-based architectures have been introduced to capture long-range dependencies within records (Wang et al., 2019; Amin-Nejad et al., 2020; Theodorou et al., 2023; Karami et al., 2024). While these models excel at learning temporal and structural patterns, they generally treat EMRs as sequences of tokens without mechanisms to ensure clinically meaningful coherence.

LLM-based EMR Generation. With the emergence of large language models (LLMs), recent studies have explored prompting LLMs to synthesize EMRs, either by providing brief clinical descriptions or by asking the model to emulate real patient records (Litake et al., 2024; Kumichev et al., 2024; Abdel-Khalek et al., 2024; Kweon et al., 2024). While LLMs exhibit strong capabilities, direct generation often results in outputs that diverge from realistic corpus-level distributions.

3 METHOD

In this section, we present **LLM-CARe**, a stage-wise cyclic refinement framework for enhancing the quality of synthetic EMRs. Unlike ordinary free-form text, EMRs are structured medical documents whose quality must be considered at multiple levels. At the *corpus level*, the dataset as a whole should follow realistic clinical distributions. At the *section level*, each field within a record should be sufficiently informative. At the *document level*, multiple fields must remain logically consistent and clinically sound when viewed together. Details are provided in Appendix A.

To address these requirements, LLM-CARe refines EMRs in three successive stages—corpus, section, and document. As illustrated in Figure 2, each stage involves the collaboration of three agents in a cyclic loop: the *Critic*, who evaluates drafts against stage-specific objectives; the *Adviser*, who pinpoints areas for improvement and suggests strategies; and the *Reviser*, who incorporates feedback to update the drafts. This iterative process progressively improves EMRs from global distributional alignment, to field-level completeness, and finally to record-level coherence.

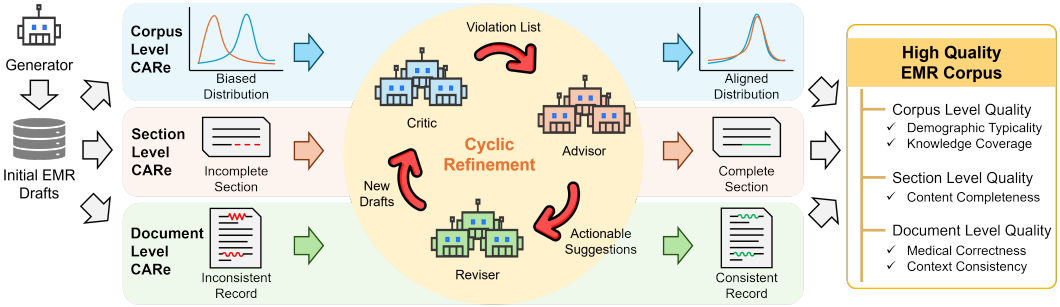

Figure 2: Overview of our proposed LLM-CARe framework for synthesizing high-quality EMRs.

3.1 INITIAL DRAFT GENERATION

The generation process begins with a *generator* agent $\mathcal{M}_{\text{generator}}$, which produces initial EMR drafts based on an input prompt x. This prompt specifies key information such as the target primary diagnosis and required EMR fields (e.g., chief complaint, history of present illness). For each prompt, the generator samples multiple drafts to form a starting draft pool:

$$\mathcal{D}^{(0)} = \{E_1^{(0)}, \ldots, E_n^{(0)}\}, \quad E_i^{(0)} \sim \mathcal{M}_{\text{generator}}(x) \tag{1}$$

where $\mathcal{D}^{(0)}$ denotes the initial draft set and each $E_i^{(0)}$ is an EMR instance. These drafts may exhibit omissions, inconsistencies, or clinically implausible details, but they provide the foundation for subsequent stage-wise refinement.

3.2 CORPUS-LEVEL CARE

At the dataset scale, high-quality synthetic EMRs must preserve realistic and representative distributions. We focus on two aspects: **Demographic Typicality**, ensuring variables such as age and

gender reflect real-world patient populations (see Figure 9 for examples), and **Knowledge Coverage**, ensuring the corpus contains both common and rare clinical conditions (with detailed dimensions listed in Table 9). These goals are addressed through corpus-level agent interactions, where the critic, adviser, and reviser collaborate to align the overall distribution.

Corpus-level Critic. At the corpus stage, the critic focuses on dataset-wide properties, capturing how well synthetic EMRs aligns with target distributions. For each attribute $c_{\text{corpus},k}$ (e.g., age, gender, or a knowledge dimension), it measures the deviation of the current corpus $\mathcal{D}^{(t)}$ from the reference distribution $\mathcal{T}_d$ derived from the training set:

$$\delta^{(t)}_{\text{corpus},k} = \mathcal{M}^{\text{corpus}}_{\text{critic}}\Big(\mathcal{D}^{(t)}, \mathcal{T}_d, c_{\text{corpus},k}\Big) \tag{2}$$

Corpus-level Adviser. The adviser interprets the critic's feedback to guide modifications at the dataset level. Based on the deviations, it identifies a subset of records $\mathcal{S}^{(t)}_k \subset \mathcal{D}^{(t)}$ whose adjustment would most effectively reduce distributional mismatch, and generates actionable feedback $F^{(t)}_{\text{corpus},k}$:

$$\mathcal{S}^{(t)}_k, F^{(t)}_{\text{corpus},k} = \mathcal{M}^{\text{corpus}}_{\text{adviser}}\Big(\mathcal{D}^{(t)}, c_{\text{corpus},k}, \delta^{(t)}_{\text{corpus},k}\Big) \tag{3}$$

Corpus-level Reviser. The reviser applies the adviser's instructions to the selected subset $\mathcal{S}^{(t)}_k$, modifying or enriching records to better match the reference distribution:

$$\mathcal{S}^{(t+1)}_k = \mathcal{M}^{\text{corpus}}_{\text{reviser}}\Big(\mathcal{S}^{(t)}_k, F^{(t)}_{\text{corpus},k}\Big) \tag{4}$$

3.3 SECTION-LEVEL CARE

At the section scale, high-quality EMRs must ensure **Content Completeness**: each field should contain the essential clinical elements expected for its type. To operationalize this, we define a set of section-specific criteria derived from clinical guidelines (see Table 6), and apply cyclic refinement with critic, adviser, and reviser agents to supplement missing information.

Section-level Critic. The critic operates for each single section. For a section $s^{(t)}_{i,m}$ in record $E^{(t)}_i$ and a criterion $c_{\text{sec},k}$ derived from clinical guidelines, it determines whether the criterion is met:

$$\delta^{(t)}_{\text{sec},i,k} = \mathcal{M}^{\text{sec}}_{\text{critic}}\Big(s^{(t)}_{i,m}, c_{\text{sec},k}\Big), \quad \delta^{(t)}_{\text{sec},i,k} \in \{0,1\} \tag{5}$$

Section-level Adviser. For unmet criteria ($\delta^{(t)}_{\text{sec},i,k} = 0$), the adviser examines the section and designs specific instructions to indicate exactly which clinical elements should be added or clarified:

$$F^{(t)}_{\text{sec},i,k} = \mathcal{M}^{\text{sec}}_{\text{adviser}}\Big(s^{(t)}_{i,m}, c_{\text{sec},k}\Big) \tag{6}$$

Section-level Reviser. Using the adviser's guidance, the reviser updates the section by incorporating the recommended elements while preserving existing content and coherence:

$$s^{(t+1)}_{i,m} = \mathcal{M}^{\text{sec}}_{\text{reviser}}\Big(s^{(t)}_{i,m}, F^{(t)}_{\text{sec},i,k}\Big) \tag{7}$$

Through this cycle, sections are progressively completed and made sufficient for their clinical role.

3.4 DOCUMENT-LEVEL CARE

At the document scale, EMRs must ensure both **Medical Correctness**—that clinical statements are valid given the diagnosis—and **Context Consistency**—that information across sections does not conflict. To make these requirements concrete, we define detailed criteria for both aspects (see Tables 7 and Table 8). To enforce them, we refine EMRs through document-level agent interactions, where the focus is on coherence across multiple sections.

Document-level Critic. The critic evaluates each record as a whole, checking constraints across sections for logical rigor and clinical plausibility. For a consistency rule $c_{\text{doc},k}$, it outputs a judgment:

$$\delta^{(t)}_{\text{doc},i,k} = \mathcal{M}^{\text{doc}}_{\text{critic}}\Big(E^{(t)}_i, c_{\text{doc},k}\Big), \quad \delta^{(t)}_{\text{doc},i,k} \in \{0,1\} \tag{8}$$

Document-level Adviser. When inconsistencies are flagged ($\delta^{(t)}_{\text{doc},i,k} = 0$), the adviser generates targeted feedback, often suggesting edits to the less influential section to restore harmony:

$$F^{(t)}_{\text{doc},i,k} = \mathcal{M}^{\text{doc}}_{\text{adviser}}\left(E^{(t)}_i, c_{\text{doc},k}\right) \tag{9}$$

Document-level Reviser. Finally, the reviser integrates this feedback to harmonize the conflicting sections and yield an updated document:

$$E^{(t+1)}_i = \mathcal{M}^{\text{doc}}_{\text{reviser}}\left(E^{(t)}_i, F^{(t)}_{\text{doc},i,k}\right) \tag{10}$$

Through this process, records are refined into coherent, consistent, and clinically valid narratives.

3.5 Stage-wise Ordering

At each stage, the critic, adviser, and reviser interact in cycles to refine the drafts according to stage-specific principles. Once the drafts have been improved at the current granularity, they are passed to the next stage, where the agent interaction continues under a different focus. The staged order is intentional: modifications at one level can influence others, so we proceed **from the most flexible to the most stringent stage**. Corpus-level refinement is relatively soft, aiming to align distributions without requiring exact matches, and is therefore performed first. While document-level refinement enforces strict logical consistency across sections, where errors could introduce serious contradictions, thus is performed last. By progressing in this order, each stage builds on the previous one while minimizing unintended conflicts. Through this staged refinement, the synthetic EMRs achieve high quality across corpus, section, and document levels.

4 Experimental Setup

Dataset To validate the effectiveness of our method, we conduct experiments on a real-world EMR dataset comprising 192k records across 302 diseases. The dataset is carefully de-identified by removing all sensitive patient information. Unlike many prior studies that focus on synthesizing a single field in isolation (e.g., chief complaint), we consider multiple fields that together capture the clinical trajectory from admission to discharge to provide a comprehensive view of each clinical episode. To ensure consistent disease distribution across subsets, we perform an 8:2 stratified split based on disease categories, maintaining proportional representation in both the training and test sets. Further details are provided in Appendix B.

Baselines We compare our method against three representative baselines: **LSTM** (Lee, 2018), an autoregressive model trained on real EMRs; **mtGAN** (Guan et al., 2021), a GAN-based method conditioned on disease labels; and **MedSyn** (Kumichev et al., 2024), which utilizes an LLM with real EMRs as in-context examples. Additionally, we include **LLM Direct**, a straightforward baseline that generates EMRs from instructions without explicit quality control. For all methods, we generate the same number of EMRs as in the test set. Since generation is conditioned on disease labels, we ensure that the disease distribution of the synthesized data exactly matches that of the real test set.

Evaluation Settings We employ two types of metrics to evaluate synthetic EMRs:

EMR quality is assessed based on the five principles introduced above. For medical correctness, content completeness, and context consistency, we adopt a "LLM-as-a-judge" approach, using a larger model to provide reliable evaluations. For demographic typicality and knowledge coverage, we compute statistical similarity to real EMRs and measure the concept coverage of clinical terms.

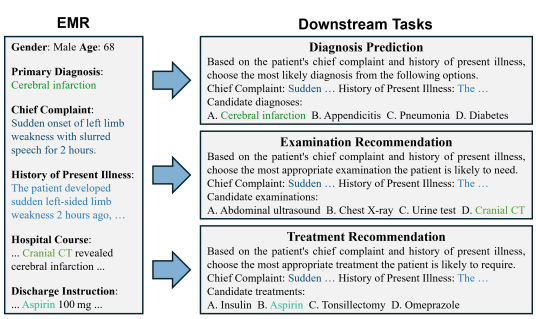

Figure 3: Construction of multiple-choice questions for three downstream tasks from an EMR.

Downstream task performance provides an practical way to evaluate the utility of synthetic

data. Following prior work, we train task-specific models on synthetic EMRs and evaluate their performance on real-world test data. This setup reflects a common use case of synthetic data in low-resource scenarios. We consider three representative tasks of high clinical relevance: diagnosis prediction, examination recommendation, and treatment recommendation. These tasks collectively span key aspects of medical decision-making. Each task is framed as a multiple-choice question, where the model predicts answers based on the chief complaint and history of present illness. Illustrative examples are shown in Figure 3.

Implementation Details We use Qwen2.5-7B-Instruct (Yang et al., 2025) as the backbone model for all agents in our framework. For EMR quality evaluation, we adopt Qwen2.5-32B-Instruct, as larger models tend to provide more reliable judgments. For downstream tasks, we fine-tune Qwen2.5-0.5B-Instruct on synthetic EMRs and evaluate performance on real test data. Detailed experimental settings are provided in Appendix C.

5 Experimental Results and Discussion

5.1 Comparison of EMR Quality

Table 1: Quality score (%) of generated EMRs across principles, where higher values indicate better performance. (*) denotes standard deviation calculated from 3 runs with different random seeds.

Type	Method	Rely on Real EMRs	Section Level	Document Level		Corpus Level	
			Content Completeness	Medical Correctness	Context Consistency	Demographic Typicality	Knowledge Coverage
Non-LLM	LSTM	✓	70.8(1.1)	65.0(0.4)	21.7(2.3)	93.3(0.6)	70.4(0.4)
	mtGAN	✓	55.8(2.9)	51.8(6.2)	21.4(4.2)	93.6(1.4)	76.3(3.7)
LLM-Based	MedSyn	✓	84.8(0.3)	95.3(0.8)	91.9(1.1)	84.1(0.9)	84.5(5.8)
	LLM Direct	×	77.1(0.1)	90.7(0.2)	87.9(0.1)	77.7(0.1)	73.9(0.2)
	LLM-CARe(ours)	×	**91.2**(0.4)	**98.6**(0.0)	**93.8**(0.1)	**96.8**(1.4)	**94.1**(0.1)

Table 1 summarizes the performance of all methods across the five defined quality principles. Our approach consistently outperforms all baselines on every metric, demonstrating its effectiveness in enhancing both the quality of individual records and the overall corpus characteristics. Among the baselines, LLM-based methods generally perform better than traditional models on section- and document-level principles. However, without our structured cyclic refinement mechanism, they sometimes fall behind traditional approaches on corpus-level principles. This suggests that **simply employing LLMs is not sufficient to guarantee comprehensive EMR quality**. In contrast, our method achieves strong and balanced performance across all levels, underscoring the benefit of integrating principle-based, stage-wise cyclic refinement into the generation process.

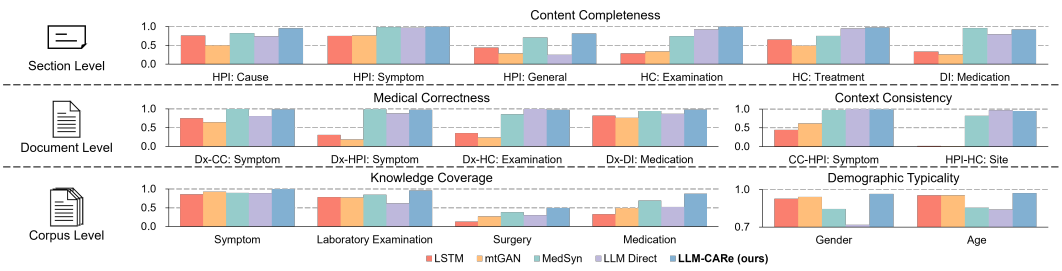

Figure 4: Detailed EMR quality evaluation across 3 levels. Abbreviations: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course, DI-Discharge Instructions, Dx-Diagnosis.

Figure 4 presents a fine-grained breakdown of EMR quality across several representative and clinically important criteria; complete results are provided in Appendix D. Our method achieves the best or competitive performance on the majority of criteria, demonstrating robustness across diverse

quality dimensions. Notably, MedSyn underperforms even the LLM Direct on some criteria, such as the completeness of the hospital course. **This suggests that real EMRs—used by MedSyn as in-context exemplars—may contain omissions that propagate into the generated records.** These findings further highlight the limitations of purely imitative approaches and emphasize the importance of explicitly modeling and enforcing quality standards during generation.

5.2 Comparison of Downstream Task Performance

Table 2: Accuracy (%) of three downstream tasks, where micro and macro are averaged across diseases. (*) denotes standard deviation calculated from 3 runs with different random seeds.

Type	Method	Rely on Real EMRs	Diagnosis Prediction		Examination Recommendation		Treatment Recommendation	
			Micro	Macro	Micro	Macro	Micro	Macro
Non-LLM	LSTM	✓	74.0(2.0)	73.1(2.0)	75.7(0.3)	76.4(0.2)	56.7(0.6)	50.0(0.7)
	mtGAN	✓	81.9(2.2)	80.9(2.5)	72.4(1.5)	73.4(1.4)	58.6(2.8)	52.9(3.0)
LLM-Based	MedSyn	✓	81.7(0.0)	81.7(0.0)	82.9(0.1)	82.2(0.1)	74.5(0.1)	71.3(0.2)
	LLM Direct	×	81.8(0.0)	81.8(0.2)	64.4(0.0)	65.4(0.0)	60.9(0.2)	59.0(0.3)
	LLM-CARe(ours)	×	**82.6**(0.3)	**82.4**(0.4)	**85.3**(0.1)	**85.2**(0.1)	**76.9**(0.3)	**74.1**(0.5)

To evaluate the utility of synthetic EMRs, we assess their effectiveness in training models for downstream tasks. As shown in Table 2, **LLM-CARe achieves the best performance across all three tasks, without using any real EMR text**. In contrast, most baselines rely on real records, either for model training or as in-context examples, which raises privacy concerns. Notably, baseline methods perform relatively well on the diagnosis task, but show larger performance gaps on examination and treatment tasks. We attribute this to their higher coverage of symptom-related knowledge (which is directly relevant to diagnosis) but limited representation of clinical concepts related to examinations, procedures, and medications—key to the latter two tasks. These results highlight the advantage of our principled framework in producing semantically rich and clinically useful synthetic records.

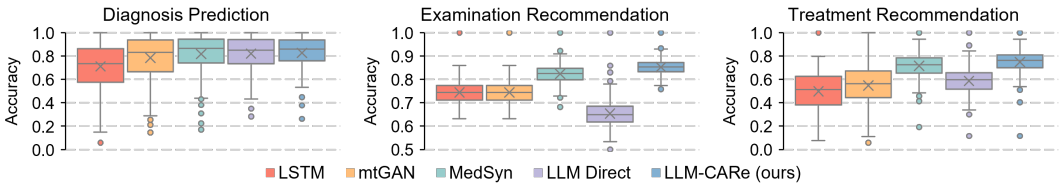

Figure 5: Accuracy distribution across diseases of different methods on three downstream tasks.

Figure 5 further illustrates the performance distribution across diseases. The box plots show that our method not only achieves higher average performance but also exhibits narrower variance across diseases. This consistency suggests that **our approach is broadly effective and robust across a wide range of clinical conditions**. In contrast, some baselines display wide performance fluctuations, indicating limited generalization to diverse disease types. These findings underscore the reliability of our method in real-world clinical settings, where robustness across varied diseases is critical.

5.3 Analysis of Performance Across Stages

Figure 6 shows how EMR quality and downstream task performance evolve through the three refinement stages. Quality dimensions improve most notably in their corresponding stages (e.g., completeness during the section stage). While some dimensions may show temporary fluctuations at other stages, the staged design—progressing from softer corpus-level constraints to stricter document-level checks—ensures that all dimensions ultimately exceed direct generation by a clear margin. For downstream tasks, examination and treatment recommendation benefit most from corpus-level refinement, since they rely on broad and diverse clinical concepts present in the training data. In contrast, diagnosis prediction depends more directly on complete histories and symptom–diagnosis alignment, thus improves primarily at the section and document stages.

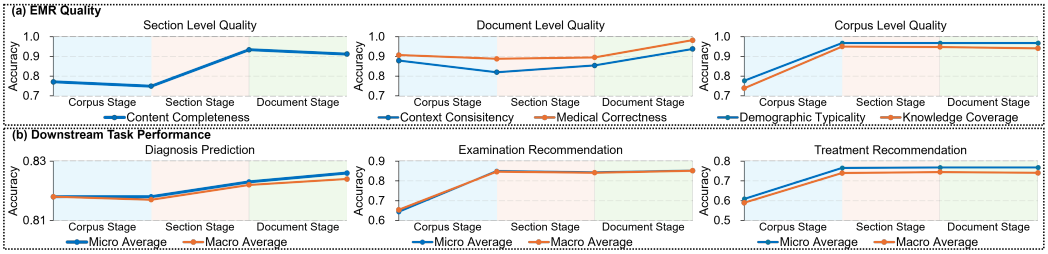

Figure 6: Trends on **(a)** EMR quality and **(b)** downstream task performance across stages.

5.4 Ablation Study of Multi-Agent Components

To assess the contribution of each agent in our framework, we conduct an ablation study by individually removing the Critic, Adviser, and Reviser agents. When the Critic is removed, the Adviser generates feedback for all quality criteria, regardless of whether they are already satisfied. Without the Adviser, the Reviser receives only high-level information about unmet criteria, without actionable suggestions. When the Reviser is removed, the system cannot update existing drafts—instead, we prompt the Generator to regenerate EMRs using all quality criteria as input.

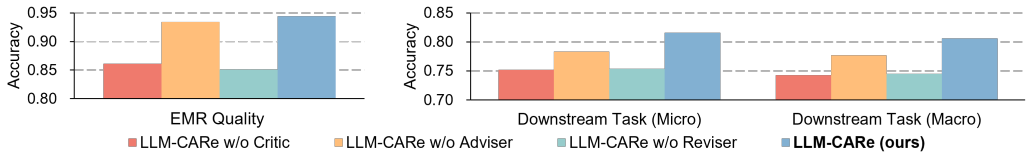

Figure 7: Impact of removing each agent on EMR quality and downstream task performance.

Figure 7 shows that removing any of the agents leads to a noticeable performance drop in both EMR quality and downstream tasks. The most significant declines occur when either the Critic or Reviser is ablated, highlighting two key insights: accurate assessment of the current draft is crucial for targeted refinement; and **large language models struggle to satisfy all quality criteria in a single generation step**, underscoring the need for cyclic refinement. Besides, removing the Adviser also results in a performance drop, suggesting that concrete, actionable feedback is more effective than abstract criterion-level input in guiding successful revisions.

5.5 Analysis of Robustness across LLM Backbone

Table 3 compares our framework with the LLM Direct baseline across three different LLM backbones: a general-purpose model (LLaMA 3.1) (Dubey et al., 2024), a medically pre-trained model (Meditron) (Chen et al., 2023), and a reasoning-oriented model (R1) (Guo et al., 2025). Without any prompt tuning or model-specific adaptation, our method consistently improves both EMR quality and downstream task performance across all backbones.

Table 3: EMR quality and downstream performance (%) across LLM backbones, averaged over principles and tasks.

Backbone	Generation Strategy	EMR Quality	Downstream Task Micro Average	Downstream Task Macro Average
Llama3.1 -8B-Instruct	LLM Direct	49.3	54.9	55.8
	LLM-CARe	**77.5**	**73.1**	**71.7**
Meditron3 -8B	LLM Direct	53.9	53.7	54.8
	LLM-CARe	**76.4**	**73.6**	**72.4**
R1-Distill -Llama-8B	LLM Direct	55.5	51.3	52.4
	LLM-CARe	**80.5**	**72.8**	**71.5**

Notably, **although Meditron is explicitly trained for medical domains, it still struggles to directly generate high-quality EMRs** and gains substantial improvements when integrated into our framework. Similarly, R1 does not significantly outperform the general model in direct generation, indicating that **internal reasoning alone is insufficient to meet the nuanced requirements of EMR**. These findings emphasize the necessity of principle-driven refinement that complements backbone capabilities and cannot be replaced by pretraining or reasoning alone.

5.6 Clinician Evaluation

To validate the reliability of using LLM as a judge, we conducted a human evaluation study. A total of 100 synthetic EMRs were sampled (20 from each of five methods) and independently assessed by four licensed clinicians, who rated completeness, consistency, and correctness for each record. As shown in Table 4, the agreement between clinicians and LLM is consistently high (Cohen's Kappa = 0.837 overall, where values exceeding 0.8 indicate near-perfect agreement), with tight confidence intervals. Inter-clinician agreement is also strong (Fleiss's Kappa = 0.947 overall), confirming that the evaluation criteria are well-defined and consistently interpretable by human experts. Together, these results demonstrate that the **LLM-based evaluation closely aligns with human judgment, supporting its validity as an efficient proxy for large-scale quality assessment**. More details are provided in Appendix E.

Table 4: Agreement between human clinicians and LLM-based evaluation on EMR quality.

Quality Level	Clinician-LLM Agreement Cohen's Kappa	Clinician-LLM Agreement Confidence Interval (95%)	Inter-Clinician Agreement Fleiss's Kappa	Inter-Clinician Agreement Confidence Interval (95%)
Section	0.866	[0.817, 0.910]	0.953	[0.923, 0.979]
Document	0.813	[0.768, 0.856]	0.941	[0.915, 0.964]
Overall	0.837	[0.804, 0.868]	0.947	[0.928, 0.964]

5.7 Case Study

Table 5 presents examples of quality issues that commonly arise when generation methods lack explicit adherence to quality standards. These cases reveal that without structured quality control, generated EMRs often exhibit missing details, medical inaccuracies, or inconsistencies.

In contrast, Figure 8 demonstrates how LLM-CARe progressively improves draft quality through refinement on different levels. This underscores the importance of stage-wise cyclic refinement in producing high-quality EMRs.

Table 5: Examples of quality issues in synthetic EMRs. Abbreviations: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course, Dx-Diagnosis.

Method	Example	Problem
LSTM	**Dx**: Uterine leiomyomas **Gender**: Male	Males do not have a uterus.
mtGAN	**HC**: Discharged after feeling stable.	No treatments are mentioned in HC.
MedSyn	**CC**: Diarrhea for 2 days. **HPI**: ... no diarrhea ...	CC mentions diarrhea, but HPI denies it.

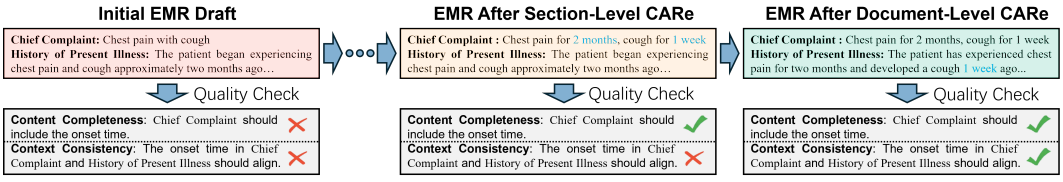

Figure 8: Illustration of quality improvements through LLM-CARe. Revisions are marked in blue.

6 Conclusion

In this work, we tackle the limitations of existing EMR synthesis methods which mimic real records without considering quality requirements. To overcome these, we propose **LLM-CARe**, a stage-wise cyclic refinement framework driven by the collaboration of **Critic**, **Adviser**, and **Reviser** agents. Instead of single-pass generation, LLM-CARe progressively enhances drafts through three dedicated stages: aligning corpus-level distributions, ensuring section-level completeness, and enforcing document-level consistency and correctness. Experiments on a large real-world dataset demonstrate that LLM-CARe substantially improves the quality of EMRs across all granularities. Moreover, models trained on the refined synthetic corpus achieve superior performance on various downstream tasks, highlighting the practical value of our approach. These results show the effectiveness of LLM-CARe in generating synthetic EMRs that are both high-quality and clinically meaningful, offering a reliable and privacy-preserving foundation for healthcare AI development.

Ethics Statement

This work adheres to the ICLR Code of Ethics. This work focuses on improving the quality of synthetic EMRs guided by clinical quality principles. The core methodology does not involve training on actual EMRs. During evaluation, a limited set of test cases was accessed within a secure, institutional data environment. These records had been fully de-identified by the hosting healthcare organization and remained within its controlled data management platform. The study did not entail any active data collection from patients or clinicians. All data usage adhered to institutional policies and was conducted under the oversight of the relevant data governance framework.

Reproducibility Statement

The collection and preprocessing of the EMR dataset are described in Section 4 and Appendix B. Experimental settings, model configurations, and evaluation protocols are detailed in Section 4 and Appendix C. The code for our experiments will be publicly released upon publication to further facilitate reproducibility.

References

Sayed Abdel-Khalek, Abeer D. Algarni, Ghada Amoudi, Salem Alkhalaf, Fahad Mohammed Alhomayani, and Shankar Kathiresan. Leveraging ai-generated content for synthetic electronic health record generation with deep learning-based diagnosis model. *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024. doi: 10.1109/TCE.2024.3415626.

Hannelore Aerts, Dipak Kalra, Carlos Sáez, Juan Manuel Ramírez-Anguita, Miguel-Angel Mayer, Juan M Garcia-Gomez, Marta Durà-Hernández, Geert Thienpont, and Pascal Coorevits. Quality of hospital electronic health record (ehr) data based on the international consortium for health outcomes measurement (ichom) in heart failure: Pilot data quality assessment study. *JMIR Medical Informatics*, 9(8), 2021. ISSN 2291-9694. doi: https://doi.org/10.2196/27842. URL https://www.sciencedirect.com/science/article/pii/S2291969421002568.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring transformer text generation for medical dataset augmentation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4699–4708, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.578/.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 12 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy142. URL https://doi.org/10.1093/jamia/ocy142.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023. URL https://arxiv.org/abs/2311.16079.

Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of heterogeneous synthetic electronic health records using gans. In *workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of

Proceedings of Machine Learning Research, pp. 286–305. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/choi17a.html.

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

Reetam Ganguli, Rishik Lad, Alice Lin, and Xiaotian Yu. Novel generative recurrent neural network framework to produce accurate, applicable, and deidentified synthetic medical data for patients with metastatic cancer. *JCO Clinical Cancer Informatics*, (7):e2200125, 2023. doi: 10.1200/CCI.22.00125. URL https://ascopubs.org/doi/abs/10.1200/CCI.22.00125. PMID: 37130342.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18 (1):173–182, 2021. doi: 10.1109/TCBB.2019.2948985.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Arun Iyengar, Ashish Kundu, and George Pallis. Healthcare informatics and privacy. *IEEE Internet Computing*, 22(2):29–31, 2018. doi: 10.1109/MIC.2018.022021660.

Hojjat Karami, David Atienza, and Anisoara Paraschiv-Ionescu. SynEHRgy: Synthesizing mixed-type structured electronic health records using decoder-only transformers. In *GenAI for Health: Potential, Trust and Policy Compliance*, 2024. URL https://openreview.net/forum?id=k4CTvnQZxx.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn: Llm-based synthetic medical text generation framework. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, pp. 215–230, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-70380-5. doi: 10.1007/978-3-031-70381-2_14. URL https://doi.org/10.1007/978-3-031-70381-2_14.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. Publicly shareable clinical large language model built on synthetic clinical notes. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5148–5168, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.305. URL https://aclanthology.org/2024.findings-acl.305/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Scott H. Lee. Natural language generation for electronic health records. *npj Digital Medicine*, 1 (1):63, Nov 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0070-0. URL https://doi.org/10.1038/s41746-018-0070-0.

Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1):98, May 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00834-7. URL https://doi.org/10.1038/s41746-023-00834-7.

Onkar Litake, Brian H Park, Jeffrey L Tully, and Rodney A Gabriel. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31(6):1404–1410, 04 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae081. URL https://doi.org/10.1093/jamia/ocae081.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1903–1911, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098088. URL https://doi.org/10.1145/3097983.3098088.

Oren Melamud and Chaitanya Shivade. Towards automatic generation of shareable synthetic clinical notes using neural language models. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann (eds.), *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 35–45, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1905. URL https://aclanthology.org/W19-1905/.

Yahia Mohamed, Xing Song, Tamara M McMahon, Suman Sahil, Meredith Zozus, Zhan Wang, Greater Plains Collaborative, and Lemuel R Waitman. Electronic health record data quality variability across a multistate clinical research network. *J Clin Transl Sci*, 7(1):e130, May 2023.

Lucy Mosquera, Khaled El Emam, Lei Ding, Vishal Sharma, Xue Hua Zhang, Samer El Kababji, Chris Carvalho, Brian Hamilton, Dan Palfrey, Linglong Kong, Bei Jiang, and Dean T. Eurich. A method for generating synthetic longitudinal health data. *BMC Medical Research Methodology*, 23(1):67, Mar 2023. ISSN 1471-2288. doi: 10.1186/s12874-023-01869-w. URL https://doi.org/10.1186/s12874-023-01869-w.

Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, 2023. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2023.100546. URL https://www.sciencedirect.com/science/article/pii/S1574013723000138.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. Smooth-gan: Towards sharp and smooth synthetic ehr data generation. In Martin Michalowski and Robert Moskovitch (eds.), *Artificial Intelligence in Medicine*, pp. 37–48, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59137-3.

Shahid Munir Shah and Rizwan Ahmed Khan. Secondary use of electronic health record: Opportunities and challenges. *IEEE Access*, 8:136947–136965, 2020. doi: 10.1109/ACCESS.2020.3011099.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1126–1133, Jul. 2019. doi: 10.1609/aaai.v33i01.33011126. URL https://ojs.aaai.org/index.php/AAAI/article/view/3905.

Yun Shen, Jiamin Yu, Jian Zhou, and Gang Hu. Twenty-five years of evolution and hurdles in electronic health records and interoperability in medical research: Comprehensive review. *Journal of Medical Internet Research*, 27, 2025. ISSN 1438-8871. doi: https://doi.org/10.2196/59024. URL https://www.sciencedirect.com/science/article/pii/S1438887125000378.

Rodrigo Tertulino, Nuno Antunes, and Higor Morais. Privacy in electronic health records: a systematic mapping study. *Journal of Public Health*, 32(3):435–454, Mar 2024. ISSN 1613-2238. doi: 10.1007/s10389-022-01795-z. URL https://doi.org/10.1007/s10389-022-01795-z.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14(1):5305, Aug 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41093-0. URL https://doi.org/10.1038/s41467-023-41093-0.

Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. Is artificial data useful for biomedical natural language processing algorithms? In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (eds.), *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 240–249, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5026. URL https://aclanthology.org/W19-5026/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1):7609, Dec 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35295-1. URL https://doi.org/10.1038/s41467-022-35295-1.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S. Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, Farhana Bandukwala, Elli Kanal, Sercan Ö. Arık, and Tomas Pfister. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine*, 6(1):141, Aug 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00888-7. URL https://doi.org/10.1038/s41746-023-00888-7.

Hongyi Yuan, Songchi Zhou, and Sheng Yu. EHRDiff : Exploring realistic EHR synthesis with diffusion models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DIGkJhGeqi.

Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. Synteg: a framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association*, 28(3):596–604, 11 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa262. URL https://doi.org/10.1093/jamia/ocaa262.

A Details of EMR Quality Principles

In this section, we provide a detailed description of each criterion corresponding to the EMR quality principles, along with representative examples.

A.1 Content Completeness

Table 6 lists the criteria used to assess content completeness, which evaluate whether each field contains all essential information.

Table 6: Description of quality criteria for content completeness, along with representative examples. Abbreviations: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course, DI-Discharge Instructions.

Criterion	Abbreviation	Positive Example	Negative Example
Chief complaint states reason for admission	CC Reason	• CC: Cough for 1 day • CC: Thyroid nodule noted for 2 months	• CC: Admitted on 2025/05/16
Chief complaint includes onset time	CC Onset	• CC: Fever for 6 days • CC: Chest pain for 1 year, worsened over past month	• CC: Dizziness accompanied by nausea • CC: Poor recent glycemic control
History of present illness describes acuity of onset	HPI Acuity	• HPI: Sudden-onset headache 1 week ago • HPI: Gradual onset of unsteady gait for 4 months	• HPI: Experienced headache over a month ago • HPI: Developed gait instability recently
History of present illness mentions possible causes	HPI Cause	• HPI: Abdominal pain after alcohol intake 1 day ago • HPI: Dizziness for 2 weeks without obvious cause	• HPI: Sudden right eye vision loss one week ago • HPI: Cough onset 3 days ago
History of present illness lists major symptoms and onset time	HPI Symptom	• HPI: Vomited 4–5 times over the past half day • HPI: Poor appetite and fatigue over past 2 weeks	• HPI: Experienced dizziness for days
History of present illness includes all general conditions	HPI General	• HPI: Normal mental status, sleep, appetite, bowel and bladder function; no significant weight change	• HPI: Appetite decreased
Hospital course includes auxiliary examinations or laboratory examinations	HC Examination	• HC: Chest CT revealed a pulmonary mass lesion • HC: Admission labs showed CRP: 12.3 mg/L	• HC: Patient underwent further examinations after admission
Hospital course includes treatment interventions	HC Treatment	• HC: Appendectomy under general anesthesia • HC: Aspirin given for antiplatelet therapy	• HC: Given pharmacological therapy
Discharge instruction includes medication dosage and usage	DI Medication	• DI: Atorvastatin 1 tablet nightly • DI: Amoxicillin 1g twice daily	• DI: Take antibiotics regularly

A.2 Medical Correctness

Table 7 outlines the criteria for medical correctness, which assess whether the clinical content aligns with the patient's diagnosis.

Table 7: Description of quality criteria for medical correctness, along with representative examples. Abbreviations: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course, DI-Discharge Instructions, Dx-Diagnosis, PD-Patient Demographics.

Criterion	Abbreviation	Positive Example	Negative Example
Diagnosis matches the patient's gender	Dx-PD Gender	• Dx: Pregnancy Gender: Female	• Dx: Pregnancy Gender: Male
Symptoms in chief complaint align with diagnosis	Dx-CC Symptom	• Dx: Pneumonia CC: Cough for 1 day	• Dx: Pneumonia CC: Knee pain for 3 days
Symptoms in history of present illness align with diagnosis	Dx-HPI Symptom	• Dx: Cerebral infarction HPI: Sudden slurred speech 1 day ago	• Dx: Acute appendicitis HPI: Sudden blurred vision 2 weeks ago
Examinations in hospital course align with diagnosis	Dx-HC Examination	• Dx: Pneumonia HC: Chest CT indicated pneumonia	• Dx: Cerebral infarction HC: Abdominal ultrasound showed appendiceal thickening
Medications in discharge instructions align with diagnosis	Dx-DI Medication	• Dx: Type 2 diabetes DI: Metformin (0.5g), one tablet twice daily	• Dx: Pneumonia DI: Insulin injection before meals

A.3 Context Consistency

Table 8 presents the criteria for context consistency, which evaluate whether information across different EMR sections is logically coherent.

Table 8: Description of quality criteria for context consistency, along with representative examples. Abbreviations: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course.

Criterion	Abbreviation	Positive Example	Negative Example
Symptoms in chief complaint are consistent with those in history of present illness	CC-HPI Symptom	• CC: Cough for 1 day HPI: ... wich cough ...	• CC: Cough for 1 day HPI: ... wichout cough ...
Onset time in chief complaint is consistent with that in history of present illness	CC-HPI Onset	• CC: Chest pain for 1 month HPI: Chest pain over past 1 month	• CC: Chest pain for 1 month HPI: Chest pain over past 2 months
Affected site in history of present illness is consistent with the site of examination or treatment in hospital course	HPI-HC Site	• HPI: Left leg pain after a fall HC: X-ray showed a fracture of the left leg.	• HPI: Left leg pain after a fall HC: X-ray showed a fracture of the right leg.

A.4 Demographic Typicality

For demographic typicality, we focus on two key patient attributes: gender and age. We evaluate whether the distributions of these attributes in the synthetic EMRs align with the target distributions. Figure 9 illustrates representative examples of gender and age distributions that are aligned with and deviate from the target distribution.

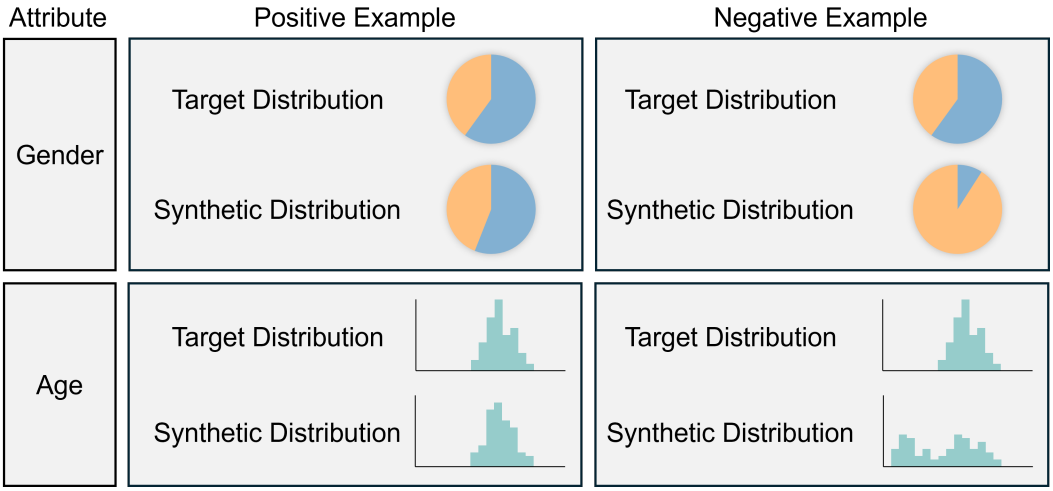

Figure 9: Examples of synthetic data distributions that are either consistent or inconsistent with the target distribution.

A.5 Knowledge Coverage

For knowledge coverage, we focus on five key categories of clinical knowledge: symptoms, auxiliary examinations, laboratory examinations, surgeries, and medications. Table 9 lists representative entities from each category.

Table 9: Knowledge categories and representative entities.

Category	Example
Symptom	Cough, Fever, Headache, Nausea, ...
Auxiliary Examination	Chest CT, Brain MRI, ECG, Abdominal Ultrasound, ...
Laboratory Examination	Complete Blood Count, Liver Function Test, C-Reactive Protein, ...
Surgery	Appendectomy, Tonsillectomy, Cataract Surgery, Cholecystectomy, ...
Medication	Aspirin, Penicillin, Metformin, Atorvastatin, ...

B Details of Dataset

B.1 Dataset Construction

We conduct our experiments on a large-scale real-world EMR dataset containing 1.82 million de-identified medical records collected from hospitals. Personally identifiable information (e.g., patient and clinician names, phone numbers, locations) had already been removed by the data provider, ensuring compliance with privacy standards. All experiments were conducted on hospital-controlled infrastructure to ensure data security and prevent risk of privacy leakage.

To ensure data quality, we first remove records that are missing critical information, such as patient age or gender, primary diagnosis, or any of the four target fields: chief complaint, history of present illness, hospital course, and discharge instruction. After this filtering step, 905k records remain.

We then apply a length-based filtering criterion to further improve data quality. During inspection, we found that overly short entries often contain placeholders or incomplete content, while excessively long entries are more likely to include unintelligible text. Therefore, we retain only records where the chief complaint is under 20 words, and each of the other fields falls within the 10 to 1,000 word range. This step yields a subset of 710k high-quality records.

Lastly, to ensure the reliability and stability of downstream evaluation, we retain only common diseases with sufficient data volume. Specifically, we exclude any diagnosis category with fewer than 500 records. This ensures that each disease has at least 100 samples in the test set after an 80/20 train-test split. Diseases with very few examples can lead to high-variance estimates and lack statistical significance in evaluation. After this final filtering step, we obtain 192k EMRs spanning 302 distinct disease categories.

B.2 Construction of Downstream Tasks

For all downstream tasks, the correct options are extracted directly from the EMR. For the diagnosis prediction task, incorrect options are randomly sampled from other diagnoses in the dataset. For the test and treatment prediction tasks, incorrect options are selected to be incompatible with the gold diagnosis: we first exclude all tests or treatments that appear in EMRs with the same diagnosis, and then randomly sample from the remaining pool. Using this approach, we construct training examples from synthetic EMRs, and evaluate model performance on questions built from real EMRs in the held-out test set.

C Experimental Details

C.1 Prompts Used for Each Agent

In this section, we list the prompts used for each agent in our LLM-CARe framework.

Generator

> Please generate an electronic medical record according to the following requirements:
> 1. The patient's primary diagnosis is: [diag].
> 2. Include only the following sections: 'Gender', 'Age', 'Primary Diagnosis', 'Chief Complaint', 'History of Present Illness', 'Hospital Course', and 'Discharge Instructions'.
> 3. Section-specific instructions:
> - The Chief Complaint should briefly describe the reason for admission.
> - The History of Present Illness should describe the onset and development of the condition in detail.
> - The Hospital Course should mention the examinations and treatments the patient received.
> - The Discharge Instructions should specify post-discharge recommendations, such as prescribed medications.
> 4. Output the result in JSON format with the structure: "Section Name": "Section Content", where each section content is a single string.

Section-Level Critic

> Below is the '[section_name]' section from an electronic medical record:
> [section]
> Please determine whether the above '[section_name]' meets the following requirement: [requirement].
> Respond in the following JSON format:
> {"Meets Requirement": true/false}

Section-Level Adviser

> Below is the '[section_name]' section from an electronic medical record:
> [section]
> This section does not meet the following requirement: [requirement]. Please provide a specific revision suggestion based on the section content, explaining how it should be modified to meet the requirement.

Respond in the following JSON format:
{"Revision Suggestion": "specific suggestion"}

Section-Level Reviser

Below is the '[section_name]' section from an electronic medical record with an issue:
[section]
The '[section_name]' section misses essential content. Please revise the record based on the following suggestion: [feedback]
Return the result in JSON format using the pattern "Section Name": "Section Content", and include only the "[section_name]" section. The content of the section should be a single string.

Document-Level Critic

Below is an electronic medical record consisting of multiple sections:
[record]
Please evaluate whether the record satisfies the following requirement: [requirement]. Identify any conflicts or implausible statements across sections.
Respond in the following JSON format:
{"Meets Requirement": true/false}

Document-Level Adviser

The following record has issues violating the requirement: [requirement]:
[record]
Please provide targeted suggestions to resolve the problem. Prioritize changes to the sections that minimally disrupt overall coherence.
Respond in the following JSON format:
{"Revision Suggestion": "specific suggestion"}

Document-Level Reviser

Below is an electronic medical record with flagged issues:
[record]
Please revise the record according to the following suggestion: [feedback].
Return the updated record in JSON format, including all sections. The content of each section should be a single string.

Corpus-Level Agents: For the corpus-level stage, we use statistical analysis tools as the critic and adviser rather than LLMs. Therefore, no natural language instructions are required for these agents; their operations are fully automated and operate on dataset-wide distributions. For the corpus-level reviser, each sample in the selected subset is modified individually, using the same type of instructions as the document-level reviser.

C.2 LLM BACKBONES

We use the following pretrained large language models in our experiments:

- **Qwen2.5** (Yang et al., 2025): Licensed under the Apache 2.0 License[1]. We use the model checkpoints available on Huggingface[2].
- **LLaMA 3.1** (Dubey et al., 2024): Licensed under the LLaMA 3.1 Community License[3]. We use the model checkpoints available on Huggingface[4].

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct/blob/main/LICENSE
[2] https://huggingface.co/Qwen
[3] https://www.llama.com/llama3_1/license/
[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

- **Meditron 3** (Chen et al., 2023): Licensed under the LLaMA 3.1 Community License[5]. We use the model checkpoints available on Huggingface[6].
- **DeepSeek-R1-Distill-Llama** (Guo et al., 2025): Licensed under the MIT License[7]. We use the model checkpoints available on Huggingface[8].

C.3 HYPERPARAMETERS

For all established baselines, we follow the original hyperparameter settings as described in their respective papers. For LLM-CARe and LLM Direct—we adopt the default generation configuration provided with each model checkpoint. For all refinement stages (corpus, section, and document), the Critic, Adviser, and Reviser agents are iterated for two cycles before proceeding to the next stage. We empirically observed that additional iterations beyond two provided negligible improvements in EMR quality and downstream task performance.

For EMR quality evaluation, we use greedy decoding to ensure deterministic outputs. For downstream tasks, we fine-tune Qwen2.5-0.5B-Instruct using the AdamW optimizer with a batch size of 16, a learning rate of 2e-5, and a cosine learning rate scheduler with 5% warmup steps. The model is fine-tuned for 3 epochs on the synthetic dataset and evaluated on the real test set.

C.4 EVALUATION DETAILS

For the LLM-based evaluation of EMR quality, we prompt the model to assess each generated EMR against the predefined criteria for medical correctness, content completeness, and context consistency. Each criterion is formulated as a binary classification task—whether a given EMR satisfies the criterion or not. The model outputs a yes/no response for each criterion per EMR, and we compute the final score by averaging over all EMRs.

For demographic typicality, we compare the distribution of demographic attributes in synthetic EMRs to those in the real dataset. For gender, we use the total variation distance (TVD) between the two distributions. For age, which is a continuous variable, we compute the Wasserstein distance. Since lower distance values indicate higher similarity, we transform the scores by computing $1 - \text{TVD}$ and $1 - \text{Wasserstein}$, respectively, so that higher values consistently reflect better quality across all metrics.

For knowledge coverage, we first extract medical entities associated with each diagnosis from the real EMRs. We then measure the proportion of these entities that appear in synthetic EMRs with the same diagnosis. To avoid the complexity and potential noise of semantic matching, we use exact string-level matching to compare entity presence.

For the downstream task evaluation, we report both macro and micro accuracy. Macro accuracy averages the model performance across all diagnoses by first computing the accuracy within each disease category, then averaging across categories. Micro accuracy, in contrast, computes the overall accuracy across all samples regardless of diagnosis. This dual evaluation provides a comprehensive view of model generalizability across frequent and less frequent disease types.

C.5 SOFTWARE

EMR generation with LLMs is conducted using vLLM (Kwon et al., 2023). PyTorch (Paszke et al., 2019) is used for training and inference of non-LLM baselines. Fine-tuning of downstream task models is performed using the Huggingface Transformers library (Wolf et al., 2020). Evaluations are also executed using vLLM.

[5] https://www.llama.com/llama3_1/license/
[6] https://huggingface.co/OpenMeditron/Meditron3-8B
[7] https://github.com/deepseek-ai/DeepSeek-R1/blob/main/LICENSE
[8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

C.6 Computational Resources

All experiments—except for EMR quality evaluation—are conducted on NVIDIA RTX 4090 GPU with 24GB of memory. EMR quality evaluation, which uses Qwen2.5-32B-Instruct, is performed on NVIDIA A100 GPU with 80GB of memory.

D Additional Experimental Results

D.1 Full Results of EMR Quality

Figure 10 provides the complete breakdown of EMR quality across all evaluated criteria, extending the representative results presented in Figure 5 of the main text. These detailed results offer a more comprehensive view of how different methods perform with respect to each quality dimension.

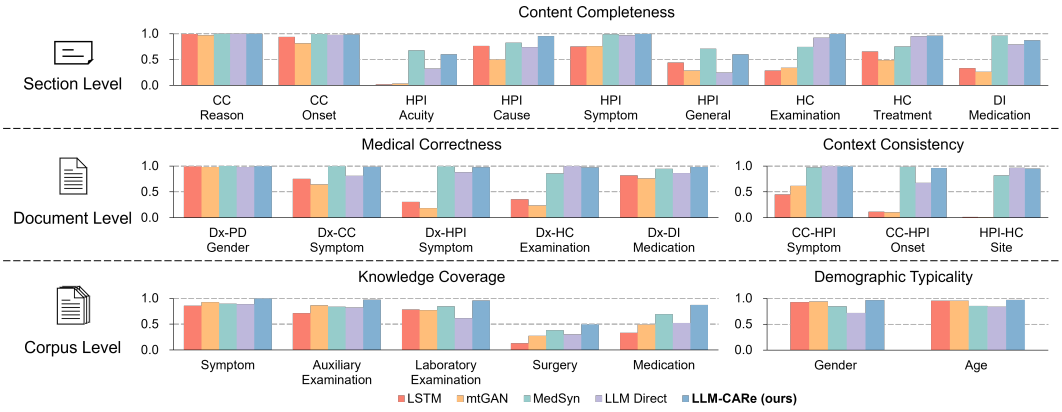

Figure 10: Fine-Grained EMR quality evaluation across three level of criteria. Abbreviations of EMR sections: CC-Chief Complaint, HPI-History of Present Illness, HC-Hospital Course, DI-Discharge Instructions, Dx-Diagnosis.

D.2 Examples of Violations on Quality Principles

To further illustrate the limitations of baseline models in generating high-quality EMRs, we provide additional examples of generated records that violate specific quality requirements.

D.2.1 Examples of Violations in Content Completeness

Method: LSTM
Generated EMR:

History of Present Illness: In 2018, the patient was discharged after recovering well from left breast cancer surgery.

Violation: The history of present illness does not include the patient's major symptoms and onset time.

Method: mtGAN
Generated EMR:

Chief Complaint: Has been eating normally for over 3 hours.

Violation: The chief complaint does not state the reason for admission.

Method: LLM Direct
Generated EMR:

History of Present Illness: The patient has a history of chronic congestive heart failure. This time, symptoms have worsened, accompanied by fatigue, chest tightness, edema, and lower limb pain. Examination revealed pulmonary moist rales and an increased heart rate.

Violation: The history of present illness does not include the patient's general condition.

Method: MedSyn
Generated EMR:

> Discharge Instruction: It is recommended to maintain a regular diet and avoid spicy and irritating foods; take medications such as omeprazole for treatment; undergo regular endoscopic follow-ups and continue follow-up visits with the gastroenterology department.

Violation: The discharge instruction does not include medication dosage and usage.

D.2.2 Examples of Violations in Medical Correctness

Method: LSTM
Generated EMR:

> Diagnosis: Rheumatoid arthritis.
> Hospital Course: After admission, relevant tests were completed, and a colon polyp repair was performed under colonoscopy.

Violation: The treatment described (colon polyp repair under colonoscopy) is not related to the diagnosis of rheumatoid arthritis, which primarily affects joints and typically involves rheumatologic evaluations and treatments, not gastrointestinal procedures.

Method: mtGAN
Generated EMR:

> Diagnosis: Synovitis of the knee joint.
> History of Present Illness: The patient developed a headache about twenty days ago and was admitted for further evaluation. Since the onset, the patient has had good mental status, appetite, and sleep, with normal bowel and bladder function.

Violation: The symptoms described in the history of present illness (headache) do not align with the diagnosis (knee joint synovitis), which typically presents with joint pain, swelling, or restricted mobility, not neurological symptoms like headache.

Method: LLM Direct
Generated EMR:

> Diagnosis: Iron Deficiency Anemia.
> Discharge Instruction: 1.Continue oral ferrous succinate 60 mg, three times daily. 2.Mecobalamin tablets 0.5 mg, once daily, to promote DNA synthesis. 3.Glimepiride tablets 2 mg, once daily, to control blood glucose.

Violation: The prescribed medications include agents that are not indicated for iron deficiency anemia, such as mecobalamin and glimepiride, which lack supporting diagnoses.

Method: MedSyn
Generated EMR:

> Diagnosis: Pulmonary embolism.
> Chief Complaint: Sudden onset of right-sided limb weakness with speech difficulty for 2 hours.

Violation: The chief complaint describes neurological symptoms suggestive of a stroke, which are not typically consistent with a diagnosis of pulmonary embolism.

D.2.3 Examples of Violations in Context Consistency

Method: LSTM
Generated EMR:

History of Present Illness: The patient experienced left knee joint pain without obvious cause one month ago and was admitted to our department for further diagnosis and treatment. Since onset, the patient has had normal mental status, appetite, sleep, and bowel and bladder function.
Hospital Course: After admission, the patient underwent relevant examinations. On 2018-03-11, a laparoscopic appendectomy was performed under general anesthesia. Postoperative treatment included anti-inflammatory therapy, fluid replacement, and symptomatic management.

Violation: The site of illness described in the history of present illness (left knee pain) is inconsistent with the site of treatment in the hospital course (appendectomy), indicating a mismatch in affected anatomical locations.

Method: mtGAN
Generated EMR:

Chief Complaint: Low back pain for 4 years, worsened over the past 5 months.
History of Present Illness: The patient experienced radiating pain in the lower limbs without obvious cause 2 years ago.

Violation: There is an inconsistency between the chief complaint and the history of present illness in both symptoms and timing. The chief complaint mentions low back pain for 4 years with recent worsening, while the history of present illness only describes lower limb radiating pain starting 2 years ago, with no mention of low back pain.

Method: LLM Direct
Generated EMR:

Chief Complaint: Sudden hearing loss with decreased hearing in the left ear, symptoms present for less than 1 day.
History of Present Illness: Two days ago, the patient experienced a sudden and significant decrease in hearing in the left ear without an obvious trigger. The hearing loss progressed to complete deafness. No associated symptoms such as vertigo, dizziness, tinnitus, ear discharge, chills, or fever were reported. The condition worsened in a stepwise manner without prior specialist treatment. The patient visited the ENT department one day before admission and was advised to be hospitalized.

Violation: The time of symptom onset stated in the chief complaint ("less than 1 day") is inconsistent with the history of present illness ("2 days ago").

Method: MedSyn
Generated EMR:

Chief Complaint: Sudden onset of chest tightness and chest pain accompanied by cold sweats for 4 hours.
History of Present Illness: Since onset, the patient has not experienced obvious fever, cough, or sputum production. Mental state is poor; appetite and sleep are average; bowel and bladder functions are normal.

Violation: The symptoms mentioned in the chief complaint are not addressed in the history of present illness.

E Details of Human Evaluation

To verify the reliability of our evaluation, we asked licensed clinicians to assess the quality of synthetic EMRs. Clinicians were instructed to read each synthetic record carefully and then answer several yes/no questions regarding **completeness**, **consistency**, and **correctness**. The questions were designed to be simple binary judgments to ensure reproducibility. The detailed labeling instructions are as follows:

Please review the synthetic EMR text shown below.

Synthetic EMR:
Gender: Male
Age: 45 years old
Primary Diagnosis: Pneumonia
Chief Complaint: Fever and cough for 3 days ...
History of Present Illness: Patient developed fever three days ago, accompanied by cough and mild chest pain ...
...

Based on the above synthetic EMR, please answer the following questions. For each question, mark your judgment in the blank: `Yes` if the requirement is satisfied, `No` otherwise.

Completeness

- Does the history of present illness include major symptoms? Answer: ____ (`Yes/No`)
- ...

Consistency

- Are the symptoms in the chief complaint consistent with those in the history of present illness? Answer: ____ (`Yes/No`)
- ...

Correctness

- Is the patient's sex valid given the diagnosis? Answer: ____ (`Yes/No`)
- ...

Confidence interval estimation: To quantify agreement, we report Cohen's Kappa and Fleiss's Kappa with 95% confidence intervals. The intervals were computed using a non-parametric bootstrap procedure with 10,000 resamples, which provides uncertainty estimates without assuming normality of the statistics.

F USE OF LLMS

In preparing this manuscript, we used LLM solely as an assistive tool for text refinement, including grammar correction, and language polishing. The research ideas, experimental design, implementation, and analysis were entirely conceived and executed by the authors.