EMRGUIDE-ZERO: SYNTHESIZING ELECTRONIC MEDICAL RECORDS VIA DUAL-LLM ADVERSARIAL GUIDELINE INDUCTION FROM SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic medical records (EMRs) are a valuable source for clinical research, yet privacy restrictions make them difficult to access and share. Synthetic EMRs offer a practical alternative, and recent methods steer large language models (LLMs) with manually summarized synthesis guidelines, often alongside real examples. Because the generator follows what these guidelines specify, fidelity hinges on them, yet manual summaries leave many context-specific requirements unstated and omit the conditions under which others apply. To address these limitations, we introduce EMRGuide-Zero, a dual-LLM adversarial framework that induces fine-grained synthesis guidelines from samples with zero manual guideline authoring. At each round, an Inducer LLM contrasts real and synthetic records to propose candidate guidelines, which are tested on broader samples to filter out local coincidences. Each guideline also specifies where it applies, allowing a Generator LLM to impose it only on the cases it concerns. The validated guidelines are incorporated into subsequent generation rounds, where new contrasts expose requirements that earlier rounds missed. Experiments across multiple EMR corpora and LLM backbones show that EMRGuide-Zero discovers fine-grained, clinically meaningful guidelines and improves fidelity and downstream utility while keeping privacy risk close to generation without direct use of real record text. These results establish guideline induction from samples as an effective alternative to manual authoring for synthesizing information-rich EMRs across clinical contexts.

1 INTRODUCTION

Electronic medical records (EMRs) are a valuable foundation for clinical research and medical artificial intelligence (Wornow et al., 2023; Shah & Khan, 2020), but privacy, ethical, and institutional safeguards substantially restrict their access and sharing (Iyengar et al., 2018; Sarkar et al., 2024). Synthetic EMRs offer a way to reproduce useful clinical patterns without releasing the original records (Yan et al., 2022; Murtaza et al., 2023), and large language models (LLMs), with their strong language understanding and long-form generation, have become a natural foundation for synthesizing realistic and detailed clinical narratives.

Recent methods steer LLM synthesis with manually summarized guidelines, such as disease-related entities drawn from knowledge bases or requirements taken from documentation standards, often complemented by examples from real records (Kumichev et al., 2024; Xu et al., 2024; Lee et al., 2026; Ning et al., 2026). Such guidelines have improved the quality and downstream utility of synthetic EMRs by telling the generator what to include and which standards to meet. The same mechanism makes the guidelines decisive for fidelity: the generator follows what they specify and otherwise relies on its own defaults, so the resulting records depend on how completely and precisely the guidelines capture the requirements that real records satisfy.

In practice, we observe two kinds of errors in records synthesized under manually summarized guidelines (Figure 1). The first reflects *incomplete coverage*. Documentation requirements differ across diseases, record sections, and patient situations, and a finite set of manually summarized guidelines can address only part of them, leaving other requirements unstated. The second reflects *insufficient granularity*. To remain broadly usable, manually summarized guidelines tend to be stated

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

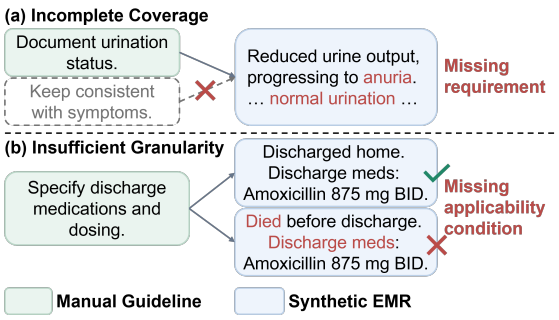


Figure 1: Errors under manual guidelines. **(a)** Incomplete coverage: no guideline keeps urination consistent with symptoms. **(b)** Insufficient granularity: a general medication guideline is applied to a deceased patient.

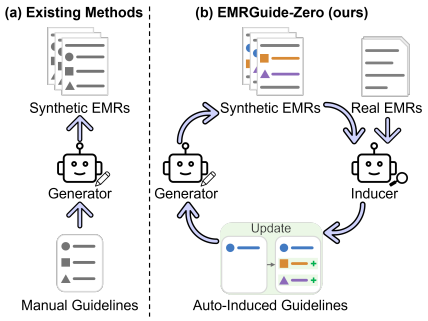


Figure 2: Fixed manual guidelines versus EMRGuide-Zero, which iteratively induces guidelines from real–synthetic differences to guide generation.

in general terms without the conditions under which they hold, so a guideline suited to most patients may be applied to cases it does not fit. Involving more experts does not readily resolve either problem, since the relevant situations are numerous and vary across corpora, and specifying each requirement down to its applicable cases demands effort that grows with every new setting. These observations motivate inducing fine-grained synthesis guidelines, together with their applicability conditions, automatically from record samples rather than summarizing them by hand.

To this end, we introduce **EMRGuide-Zero**, a dual-LLM adversarial framework that induces synthesis guidelines from samples without manual authoring (Figure 2). Rather than enumerating requirements in advance, it locates them where current generation falls short, in the differences between real and synthetic records. An Inducer LLM proposes candidate guidelines from these differences and validates them on broader samples to exclude incidental details, and a Generator LLM synthesizes the next corpus under the accepted guidelines. As salient gaps close, new contrasts expose finer ones, so the guideline bank grows with the generator's remaining errors. Each guideline also specifies where it applies, either by a condition on the record or, for content that no single record must contain, by its frequency and co-occurrence across records, so that the Generator imposes each guideline only on the cases it concerns rather than on every record.

Experiments on two EMR corpora with three LLM backbones show that EMRGuide-Zero improves fidelity and semantic coverage in most settings and achieves the best diagnosis-prediction accuracy in every setting, with lower privacy risk than methods conditioned on real examples. Guidelines transferred to another generator perform on par with its native induction. Our contributions are:

- We propose automatically inducing EMR synthesis guidelines from record samples, capturing fine-grained requirements and applicability conditions that are difficult to summarize manually across diverse diseases and patient situations.
- We develop EMRGuide-Zero, a dual-LLM adversarial framework in which an Inducer induces guidelines from real–synthetic differences and a Generator uses them to synthesize the next round of records, whose remaining gaps drive further induction.
- Experiments across corpora and backbones show that the induced guidelines improve fidelity and downstream utility, transfer across generators, and keep privacy risk close to that of generation without direct use of real record text.

2 Related Work

Neural EMR Generation. Adversarial models synthesize structured or textual EMRs (Choi et al., 2017; Baowaly et al., 2019; Guan et al., 2021; Yoon et al., 2023); autoregressive and sequential models capture longitudinal events or clinical narratives (Lee, 2018; Melamud & Shivade, 2019; Zhang et al., 2021; Theodorou et al., 2023); and diffusion models learn record distributions (Naseer et al., 2023; Yuan et al., 2024). These methods established fidelity, privacy, and utility as central objectives but have limited instruction-following compared with current LLM-based synthesis.

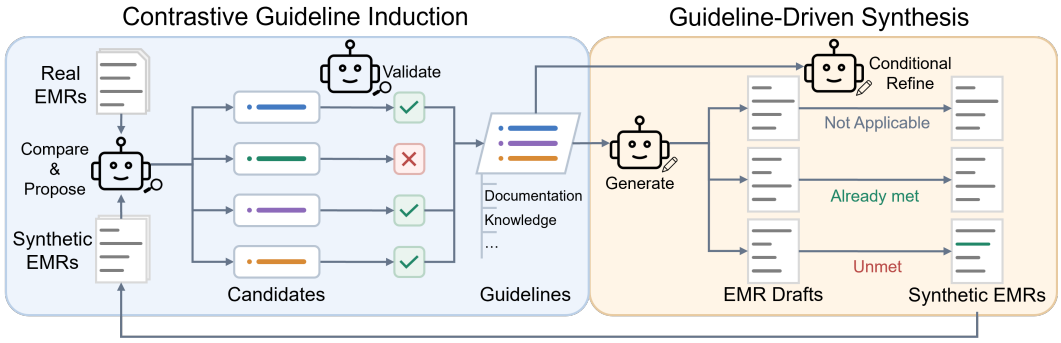


Figure 3: Overview of EMRGuide-Zero. In guideline induction, the Inducer proposes guidelines from real–synthetic differences and validates them on broader samples. In guideline-driven synthesis, the Generator plans record content with content guidelines and refines drafts with applicable documentation guidelines; the new synthetic records drive the next round of induction.

LLM-Based Clinical-Text Generation. Strong language understanding, instruction following, and long-form generation have made LLMs a common foundation for synthetic clinical text. Recent methods control them with supplied guidelines. MedSyn combines disease-specific symptoms sampled from a medical knowledge graph with examples from real records (Kumichev et al., 2024), while ClinGen derives clinical topics and writing styles from external domain knowledge (Xu et al., 2024). Expert-informed prompts and feedback loops generate narratives from case descriptions (Lee et al., 2026). LLM-CARe organizes predefined clinical-quality principles together with aggregate demographic and marginal entity-frequency targets from real records into corpus-, section-, and document-level Critic–Adviser–Reviser cycles (Ning et al., 2026). These approaches show that supplied guidelines can improve controllability, but their content and applicability are determined before generation. Marginal targets do not express relationships among entities, while general criteria or record examples may not fit every case. These limitations motivate automatic guideline induction from samples to capture corpus-specific requirements without manual authoring.

3 METHOD

Let $\mathcal{D}_{\text{real}}$ denote a real EMR corpus and $e \in \mathcal{D}_{\text{real}}$ an individual record. We seek a synthetic corpus $\mathcal{D}_{\text{syn}}$, with $\tilde{e} \in \mathcal{D}_{\text{syn}}$ denoting a synthetic record, that can substitute for the real corpus in downstream use. Instead of relying on manually summarized guidelines, EMRGuide-Zero learns a bank of synthesis guidelines $\mathcal{G}$ from the records themselves. Two LLM roles, which can share a backbone, divide this work (Figure 3): the *Inducer* turns what current synthetic records fail to reproduce into validated guidelines, and the *Generator* synthesizes records under these guidelines. All steps operate within a diagnosis group, whose index we omit for brevity.

3.1 CONTRASTIVE GUIDELINE INDUCTION

Describing real records in isolation would yield many properties that the generator already reproduces. The requirements worth adding are those that current generation misses, and these appear as systematic differences between real and synthetic records. Induction therefore has two steps: the Inducer proposes candidate guidelines from such differences, and each candidate must then be supported by broader evidence before it can guide generation.

Proposing candidates from contrasts. The Inducer compares batches of real and synthetic records rather than single pairs, since any two records differ in many incidental ways, whereas differences that recur across a batch indicate systematic gaps. Each proposal must be checkable on other records and, to avoid the insufficient granularity of general guidelines, must also specify where it applies. How applicability can be specified depends on whether it can be decided from an individual record. For much clinical content, such as a symptom, a test result, or a treatment, a single record offers no basis for requiring its presence or absence: a symptom typical of a disease appears in only some patients, so what matters is how often the content appears and alongside what. Such a proposal, which

we call a *content guideline*, is paired with an executable detector γ_j, where $\gamma_j(e) = 1$ indicates that record e contains the described content, and its applicability is captured by how the content is distributed across records. Detectors are short pattern-matching programs, making their evaluation on every real record inexpensive and reproducible. Other requirements govern how information is documented and must hold in every record where a given situation arises, such as explaining how an observation supports a subsequent decision. Such a proposal, a *documentation guideline*, is written as a pair (C_k, T_k), where the condition $C_k(e) = 1$ indicates that the requirement applies to e and the target $T_k(e) = 1$ indicates that e satisfies the requirement.

Validating candidates. A proposal derived from a few records may describe a coincidence, so each candidate is validated according to how its applicability is specified. A content guideline must recur in real records, and because its applicability is expressed through its distribution across patients, that distribution must be measured rather than estimated. We therefore replay its detector over the real corpus rather than relying on the LLM's impression, and measure how often content guidelines j and ℓ hold together:

$$\pi_{j\ell} = \frac{1}{|\mathcal{D}_{\text{real}}|} \sum_{e \in \mathcal{D}_{\text{real}}} \gamma_j(e)\gamma_\ell(e). \tag{1}$$

The diagonal π_{jj} indicates how often the content appears, and the off-diagonal terms indicate which contents tend to appear together. Recurring candidates and their measurements form the content guideline set $\mathcal{G}_{\text{cont}}$, grounding generation in source evidence. A documentation guideline must instead be met consistently by applicable real records while remaining unmet by current generation. Let $\mathcal{V}_{\text{real}}$ denote real records excluded from the proposal examples and $\mathcal{V}_{\text{syn}}$ synthetic records from the current round. For $X \in \{\text{real}, \text{syn}\}$, the target satisfaction rate among applicable records is

$$\hat{\rho}_k^X = \frac{\sum_{e \in \mathcal{V}_X} C_k(e)T_k(e)}{\sum_{e \in \mathcal{V}_X} C_k(e)}. \tag{2}$$

Restricting both rates to applicable records keeps a requirement from being judged on cases it does not concern. With an admission threshold $\tau \in [0, 1]$, a candidate enters the documentation guideline set $\mathcal{G}_{\text{doc}}$ only if $\hat{\rho}_k^{\text{real}} > \tau$, confirming that real records consistently meet it, and $\hat{\rho}_k^{\text{real}} > \hat{\rho}_k^{\text{syn}}$, confirming that current generation still falls short. Together, the two conditions retain only requirements that have recurring source support and can still improve the current generator.

3.2 Guideline-Driven Synthesis

Validated guidelines tell the Generator what to reproduce, but placing all of them in every prompt would give every patient the same common content and apply requirements to records they do not concern. Following how their applicability is specified, the two kinds of guidelines enter generation at different points: content guidelines determine what each record contains before drafting, and documentation guidelines correct each draft only where they apply. This order ensures that documentation requirements, whose applicability depends on the content of a record, are checked only after that content is settled rather than on a draft that may still change.

Content planning. Each synthetic record should contain content plausible for a single patient, while the corpus as a whole should preserve how content varies and combines across patients. Selecting content independently can pair items that rarely occur together, whereas assigning all retained content makes records unnaturally uniform. We therefore sample each record's content jointly so that, across the synthetic corpus, content guidelines follow the measured frequencies and co-occurrences $\{\pi_{j\ell}\}$, and we match the number of planned items to its empirical distribution in $\mathcal{D}_{\text{real}}$ so that plans are as dense as real records (Appendix D). For record i, the resulting content plan $\mathcal{P}_i$, the set of indices of the content guidelines assigned to it, specifies what the record should contain before any text is written and guides LLM drafting of the initial record:

$$\tilde{e}_i^{\text{init}} = \text{Generate}(\mathcal{P}_i). \tag{3}$$

Because an LLM may omit planned content, the detectors identify the missing items $\mathcal{M}_i = \{j \in \mathcal{P}_i \mid \gamma_j(\tilde{e}_i^{\text{init}}) = 0\}$, and a targeted revision adds them:

$$\tilde{e}_i^{\text{cont}} = \text{Revise}_{\text{cont}}\left(\tilde{e}_i^{\text{init}}, \mathcal{M}_i\right). \tag{4}$$

A revision is kept only if it realizes the requested content without removing planned content already present, so that repairing one omission does not create another.

Conditional refinement. Documentation guidelines require a different form of control. Applying one outside its condition adds irrelevant content, while revising a record that already satisfies it can disturb an otherwise sound draft. We therefore select only the guidelines that apply to the draft and remain unmet, $\mathcal{U}_i = \{k \mid C_k(\tilde{e}_i^{\text{cont}}) = 1,\ T_k(\tilde{e}_i^{\text{cont}}) = 0\}$, and supply their targets to a refinement step that yields the final record:

$$\tilde{e}_i = \text{Revise}_{\text{doc}}\big(\tilde{e}_i^{\text{cont}}, \{T_k \mid k \in \mathcal{U}_i\}\big)\,. \tag{5}$$

Focusing the Generator on a few unmet targets keeps each edit local instead of asking it to reconsider the entire record. An edit is retained only if a post-check confirms that the targets are satisfied without contradicting the unchanged content, so that refinement introduces no new errors.

3.3 Adversarial Guideline Induction

A single contrast mainly exposes the most salient gaps. Once guidelines close them, the synthetic corpus changes, and finer gaps that were previously masked become visible. We therefore couple the two roles in an adversarial loop: the Inducer seeks recurring properties that distinguish real records from current synthetic ones, and the Generator attempts to eliminate these differences. This progressive expansion addresses incomplete coverage, since requirements missed by one contrast can still be found in later rounds. What is learned is the guideline bank rather than model parameters, so the acquired knowledge remains explicit, can be inspected, and can be reused by other generators without repeating the induction process on the same corpus.

Starting from an empty bank, at round t the current bank $\mathcal{G}^{(t)} = \mathcal{G}_{\text{cont}}^{(t)} \cup \mathcal{G}_{\text{doc}}^{(t)}$ produces the synthetic corpus $\mathcal{D}_{\text{syn}}^{(t)}$. Contrasting it with $\mathcal{D}_{\text{real}}$ yields newly validated content and documentation guidelines $\Delta\mathcal{G}_{\text{cont}}^{(t)}$ and $\Delta\mathcal{G}_{\text{doc}}^{(t)}$, and the bank grows as

$$\mathcal{G}^{(t+1)} = \mathcal{G}^{(t)} \cup \Delta\mathcal{G}_{\text{cont}}^{(t)} \cup \Delta\mathcal{G}_{\text{doc}}^{(t)}. \tag{6}$$

Because each round contrasts real records with the latest synthetic corpus, induction targets what current generation still misses rather than rediscovering gaps that are already closed. The loop terminates when no new guideline is validated for several consecutive rounds or a preset round limit is reached, and the final bank is used to generate the released synthetic corpus.

4 Experimental Setup

Datasets. We evaluate on two diagnosis-stratified corpora: **EMR-302**, a deidentified inpatient corpus covering 302 primary diagnoses (Ning et al., 2026), and **MIMIC-47**, a subset constructed from MIMIC-IV-Note v2.2 (Johnson et al., 2023). Following the prior protocol, we split EMR-302 80/20 within each diagnosis. For MIMIC-47, we construct a patient-disjoint 80/20 split and retain diagnosis groups containing at least 400 training and 100 test patients, yielding 47 groups. We use only the resulting training records for guideline induction and baseline construction, reserving test text for final evaluation. Dataset construction and statistics are provided in Appendix A.

Baselines. To distinguish learned guidelines from supplied ones, we compare with **Direct** generation, knowledge- and example-guided **MedSyn** (Kumichev et al., 2024), topic- and style-guided **ClinGen** (Xu et al., 2024), and fixed-criterion iterative refinement with **LLM-CARe** (Ning et al., 2026). Within each experimental cell, all methods use the same backbone, conditions, output schema, and corpus size. Details of baselines are shown in Appendix B.

Evaluation. We assess **fidelity**, or how closely synthetic records resemble real records, through entity frequency, entity co-occurrence, and semantic similarity. These capture how often clinical entities appear, how they occur together, and how closely average semantic content aligns, measured by MAE, joint TVD, and MMD^2, respectively. The entity inventory is fixed from training data, shared by all methods, and independent of the guidelines learned by EMRGuide-Zero. The same entity-matching procedure scores real and synthetic records. For **diversity**, we assess semantic coverage: whether the synthetic corpus represents the range of content in real cases, measured by real-to-synthetic nearest-neighbor distance. For **utility**, we test whether synthetic records support diagnosis prediction by training a fixed model on synthetic records and measuring its accuracy on

Table 1: Overall comparison across datasets and generation backbones. Entity frequency: MAE; entity co-occurrence: joint TVD; semantic similarity: linear-kernel MMD^2; semantic coverage: real-to-synthetic nearest-neighbor cosine distance; diagnosis prediction: accuracy on $[0, 1]$. Scores are diagnosis-macro averages; bold marks the best unrounded score for each dataset and backbone.

	EMR-302					MIMIC-47				
Method	Entity freq. ↓	Entity cooc. ↓	Semantic sim. ↓	Semantic cov. ↓	Diagnosis pred. ↑	Entity freq. ↓	Entity cooc. ↓	Semantic sim. ↓	Semantic cov. ↓	Diagnosis pred. ↑
Qwen3.5-9B										
Direct	0.293	0.468	0.187	0.239	0.821	0.287	0.448	0.283	0.325	0.807
MedSyn	0.210	0.338	0.097	0.210	0.842	0.221	0.336	0.168	0.295	0.821
ClinGen	0.238	0.380	0.123	0.214	0.817	0.216	0.329	0.183	0.297	0.824
LLM-CARe	0.193	0.320	0.135	0.220	0.826	0.251	0.381	0.228	0.311	0.785
EMRGuide-Zero	**0.159**	**0.264**	**0.076**	**0.201**	**0.854**	**0.184**	**0.303**	**0.142**	**0.281**	**0.844**
Gemma-4-12B-it										
Direct	0.299	0.479	0.252	0.270	0.820	0.320	0.505	0.346	0.346	0.789
MedSyn	0.195	0.309	0.104	0.218	0.835	0.250	0.383	0.189	0.300	0.810
ClinGen	0.199	0.317	0.116	**0.215**	0.824	0.255	0.391	0.212	0.303	0.803
LLM-CARe	0.198	0.320	0.158	0.234	0.818	0.275	0.424	0.270	0.317	0.798
EMRGuide-Zero	**0.151**	**0.249**	**0.100**	0.218	**0.846**	**0.181**	**0.295**	**0.145**	**0.277**	**0.848**
Llama-3.1-8B-Instruct										
Direct	0.300	0.481	0.196	0.265	0.809	0.297	0.464	0.300	0.337	0.803
MedSyn	0.146	0.231	**0.066**	0.210	0.852	0.206	0.313	0.132	0.279	0.835
ClinGen	**0.113**	**0.182**	0.077	0.207	0.842	0.206	0.315	0.155	0.283	0.840
LLM-CARe	0.205	0.316	0.145	0.251	0.828	0.270	0.412	0.228	0.319	0.808
EMRGuide-Zero	0.121	0.195	0.068	**0.207**	**0.859**	**0.188**	**0.304**	**0.128**	**0.274**	**0.848**

real cases. All methods share diagnosis quotas, the downstream probe, training procedure, and real test questions. Implementation and aggregation details appear in Appendix C.

Implementation details. We use Qwen3.5-9B (Qwen Team, 2026), Gemma-4-12B-it (Gemma Team et al., 2026), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the backbone shared by the Inducer and Generator in each experimental setting. For downstream diagnosis prediction, we fine-tune Qwen3.5-0.8B on each synthetic corpus with the same training recipe. Backbone and guideline induction settings appear in Appendix D.

5 Experimental Results

5.1 Overall Comparison with Baselines

Learned guidelines better preserve clinical content and its combinations. EMRGuide-Zero leads entity-frequency and co-occurrence alignment in most settings (Table 1). Closer entity frequencies reduce over- and under-representation of findings. Better co-occurrence alignment brings combinations of findings within records closer to real practice. This addresses the insufficient granularity of fixed guidelines: knowing that a finding is relevant to a diagnosis does not establish how often, or alongside which other findings, it should appear.

Better coverage reflects diversity grounded in real cases. The semantic gains, most consistent on MIMIC-47, extend beyond the fixed entity inventory. Improved coverage means that varied real cases find closer counterparts in an equally sized synthetic corpus. Together with better average semantic alignment, this suggests broader representation of real clinical presentations. Expanded guidance thus supports useful diversity, reducing concentration on a narrow set of typical records rather than merely increasing variation in wording.

The preserved content remains useful for diagnosis. EMRGuide-Zero achieves the best diagnosis accuracy in every evaluated setting, improving by up to 3.8 percentage points over the strongest baseline. Successful transfer from synthetic training records to real test cases indicates that gener-

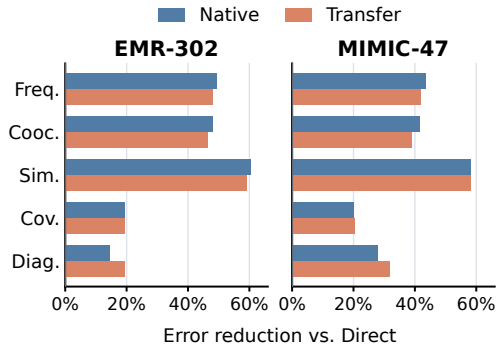


Figure 4: Guideline transfer to Gemma-4-12B-it. Bars show error reduction relative to Direct; Native uses Gemma-induced guidelines, and Transfer uses frozen Qwen-induced ones. Metrics are abbreviated as in Table 2.

Table 2: Ablation with Qwen3.5-9B. +Content adds content planning, and +Documentation further adds conditional refinement; bold marks the best score for each dataset. Freq., Cooc.: entity frequency and co-occurrence; Sim., Cov.: semantic similarity and coverage; Diag.: diagnosis-prediction accuracy.

Setting	Freq.↓	Cooc.↓	Sim.↓	Cov.↓	Diag.↑
EMR-302					
Direct	0.293	0.468	0.187	0.239	0.821
+Content	**0.157**	**0.261**	0.077	0.208	0.850
+Documentation	0.159	0.264	**0.076**	**0.201**	**0.854**
MIMIC-47					
Direct	0.287	0.448	0.283	0.325	0.807
+Content	0.186	0.307	0.155	0.290	0.839
+Documentation	**0.184**	**0.303**	**0.142**	**0.281**	**0.844**

ation preserves clinically discriminative admission information. The gains therefore translate into useful training data beyond closer distributional resemblance.

5.2 Cross-Backbone Guideline Transfer

We transfer the frozen guideline bank induced with Qwen3.5-9B to Gemma-4-12B-it and compare it with Direct and with Gemma generation under its own induced guidelines. All conditions use the Gemma generator, the same diagnosis quotas, and the five evaluation metrics. Figure 4 reports each metric's error reduction relative to Direct, using one minus accuracy as diagnosis error.

Transferred guidelines perform on par with native induction. On both datasets, Transfer stays within 2.6 percentage points of Native in error reduction on the four distributional metrics and yields slightly higher diagnosis accuracy (0.855 vs. 0.846 on EMR-302 and 0.856 vs. 0.848 on MIMIC-47), while both remain far ahead of Direct. The guidelines therefore remain useful beyond the inducing model's particular errors, suggesting that they capture recurring characteristics of the source records.

This makes the guideline bank a reusable resource. When a new model is adopted for the same clinical corpus, a previously induced bank can guide its generation directly, avoiding the cost of repeating induction and allowing guidelines accumulated with one model to benefit later generators.

5.3 Ablation Study

Table 2 examines the contribution of each guideline use with Qwen3.5-9B on both datasets. Starting from Direct, +Content adds content planning with content guidelines, and +Documentation further refines each draft with applicable documentation guidelines.

Content planning establishes distributional alignment. Adding content guidelines improves all five metrics on both datasets, with the largest gains in entity frequency, co-occurrence, and semantic similarity. Planning each record's content to follow measured frequencies and co-occurrences brings generated records closer to real clinical content while improving diagnosis transfer. This addresses a limitation of disease-related concept lists, which name relevant findings but do not specify how often they should occur or which combinations belong in a case.

Conditional refinement adds case-specific detail. Documentation guidelines further improve semantic coverage and diagnosis accuracy on both datasets and all five metrics on MIMIC-47, where semantic similarity improves from 0.155 to 0.142. On EMR-302, entity metrics stay within 0.003 of +Content, so refinement largely preserves the planned content. Better coverage indicates that varied real cases find closer synthetic counterparts: by applying each requirement only where its condition holds, refinement adds detail specific to each case rather than uniform content.

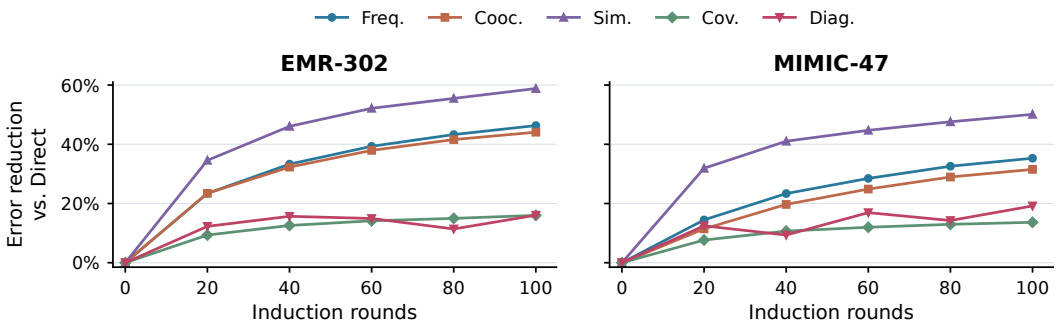


Figure 5: EMRGuide-Zero performance across induction rounds. Each curve reports a metric's error reduction relative to Direct, with metrics abbreviated as in Table 2; higher is better.

5.4 Performance across Induction Rounds

Figure 5 compares records generated using guideline banks obtained after different numbers of induction rounds, with the same generation and evaluation procedures as in the main experiment. Direct provides the round-zero reference for each metric.

Iteration expands guidelines to address remaining generation gaps. Across both datasets, continued induction brings overall gains in fidelity, coverage, and downstream utility, with benefits extending beyond the initial guideline bank. As guidelines change the generated records, new contrasts expose requirements that current generation still fails to satisfy. Validating these requirements on real and synthetic records before adding them to the bank makes subsequent generation responsive to the remaining gaps. The overall improvements support this feedback process as a way to progressively approximate the content and variation of real clinical records, extending guidance beyond the coverage of requirements fixed before generation.

5.5 Stratified Performance

We stratify performance by sex, age, and training-set disease frequency. Scores average relative error reductions over five metrics and three backbones, using Direct as reference and one minus accuracy as diagnosis error. Figure 6 shows both datasets on the same radial scale.

The benefit is distributed across clinical contexts. EMRGuide-Zero has the largest composite gain in every evaluated stratum on both datasets, so the overall advantage does not depend solely on a particular demographic group or frequent diagnoses. Benefits in less frequent diagnosis groups suggest coverage beyond common clinical scenarios; those across age and sex groups support relevance to varied patient presentations. These consistent aggregate gains indicate that learned guidelines accommodate different clinical content needs across the evaluated strata.

5.6 Privacy Evaluation

We evaluate nearest-training-record character 5-gram similarity and diagnosis-matched membership inference. Figure 7 reports similarity p99 and the absolute membership AUC gap from chance, averaged over three backbones; lower values indicate less source-text reuse and weaker membership signals. Direct and LLM-CARe generate without real-record text, whereas MedSyn and ClinGen condition on real examples. Appendix C.4 gives the evaluation details.

Source-informed quality need not require example-level exposure. MedSyn and ClinGen achieve competitive quality in some settings, but their greater source-text similarity highlights the risk of reproducing individual details from examples. EMRGuide-Zero learns from real records while keeping its overall privacy scores broadly comparable to those of Direct and LLM-CARe and below those of example-conditioned methods, with markedly less source-text reuse. It thus combines the stronger overall quality established in the preceding experiments with low exposure signals, showing the value of learning from real records through reusable guidelines.

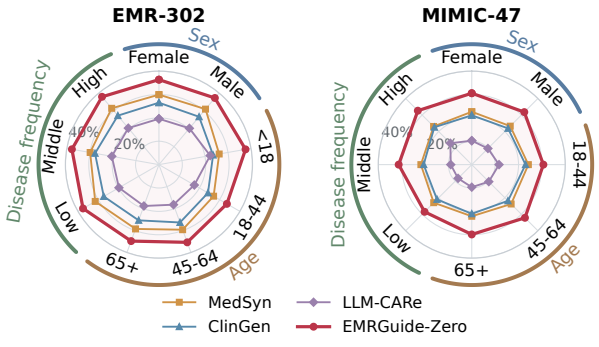


Figure 6: Stratified performance on both datasets. Scores average relative error reduction over five metrics and three backbones; higher is better. Outer arcs mark sex, age, and disease-frequency groups on a shared radial scale. Direct is the zero reference and has no separate curve.

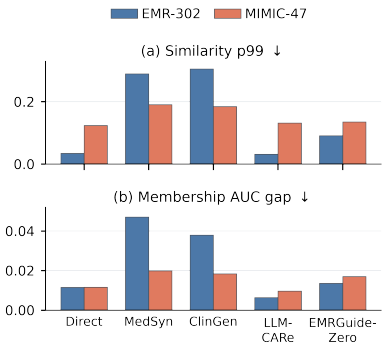


Figure 7: Privacy over three backbones. Similarity p99 measures text reuse; membership AUC gap measures distance from chance. Lower scores mean less exposure.

Requiring support beyond the proposal examples makes isolated patient details less likely to enter the guideline bank. The generator receives recurring requirements and measured content distributions instead of source passages, retaining useful clinical structure while limiting opportunities to reproduce individual details. The observed privacy–utility balance is consistent with this validation mechanism, supporting guideline induction as a way to preserve source structure through reusable guidance without passing individual patient records to the generator.

5.7 Case Study

Coverage beyond entity lists and predefined checks. Entity-based guidelines specify which clinical terms to include, but not how their information should be expressed. EMRGuide-Zero discovers a pattern linking a sodium test to its numeric result, capturing a test–value relationship beyond the test name. Similarly, requiring a chief complaint to be complete and consistent with other fields does not ensure that it clearly conveys the main reason for the visit. An induced documentation guideline requires it to state that reason concisely, without detailed descriptions of accompanying symptoms. This extends guidance to the selection and organization of clinical information.

Applicability through co-occurrence and explicit conditions. Sampling clinical entities independently does not capture which ones tend to appear together. EMRGuide-Zero discovers that two blood tests, alkaline phosphatase and total bilirubin, are reported together in the source records more often than independent sampling would imply, informing content planning. Applicability also depends on the record: an induced guideline requires discharge instructions that warn of possible adverse reactions to specify what action to take or where to seek help. The warning triggers this requirement, so the response pathway is not imposed on every record. This links a safety warning to an actionable response while restricting the requirement to relevant cases.

6 Conclusion

We introduced EMRGuide-Zero, a dual-LLM adversarial framework that induces synthesis guidelines from samples without manual authoring. Its Inducer identifies recurring real–synthetic differences and validates candidate guidelines on broader samples, while its Generator uses the resulting guidance to synthesize records. Guidelines validated across the real corpus plan what each record contains, and guidelines whose applicability can be judged from an individual record refine each draft only where their conditions hold. Experiments across multiple EMR corpora and LLM backbones show that these fine-grained guidelines improve fidelity and downstream utility, transfer across generators, and reduce the record-reuse signals associated with example-conditioned methods. These results establish automatic guideline induction as an effective approach to capturing clinical structure and synthesizing information-rich EMRs across varied clinical contexts.

AI USE STATEMENT

LLMs are an integral component of the proposed method: they induce synthesis guidelines and generate the synthetic records studied in this paper, as described in Section 3 and Appendix D. In addition, we used generative AI tools to assist with writing and editing the manuscript and with writing experiment and plotting code. We have reviewed AI-generated text and code, and we take full responsibility for the content of this paper.

ETHICS STATEMENT

All experiments use existing deidentified EMR datasets. We obtained each dataset through its provider's access procedure and followed its data use requirements. All data processing, including LLM inference, was performed on local servers, and no patient records were uploaded to external services. We will not distribute any data derived from these datasets. The synthetic records are intended solely for research purposes and not for real clinical decision-making.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix D describe the proposed method and its implementation. Section 4 and Appendices A–C describe the datasets, baselines, evaluation metrics, and experimental settings. We will publicly release all code for data preprocessing, experiments, and evaluation, so that researchers who obtain the datasets from their providers can reproduce our pipeline.

REFERENCES

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, 2019. doi: 10.1093/jamia/ocy142.

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 *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 286–305. PMLR, 2017. URL https://proceedings.mlr.press/v68/choi17a.html.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. doi: 10.1109/TIT.1968.1054142.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. doi: 10.48550/arXiv.2607.02770.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

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.

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.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, January 2023. doi: 10.13026/1n74-ne17. URL https://doi.org/10.13026/1n74-ne17. Version 2.2.

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*, pp. 215–230, 2024. doi: 10.1007/978-3-031-70381-2_14.

David Seung U Lee, Seeun Park, Seoyoon Jang, Sunyoung Lee, Chaeyoung Chang, Sungwook Choi, and Howard Lee. Generation of bilingual synthetic clinical notes for realistic data augmentation. In *Proceedings of the Seventh Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 834–879. PMLR, 2026. URL https://proceedings.mlr.press/v333/lee26b.html.

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

Oren Melamud and Chaitanya Shivade. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 35–45. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1905.

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. doi: 10.1016/j.cosrev.2023.100546.

Ahmed Ammar Naseer, Benjamin Walker, Christopher Landon, Andrew Ambrosy, Marat Fudim, Nicholas Wysham, Botros Toro, Sumanth Swaminathan, and Terry Lyons. ScoEHR: Generating synthetic electronic health records using continuous-time diffusion models. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pp. 489–508. PMLR, 2023. URL https://proceedings.mlr.press/v219/naseer23a.html.

Chen Ning, Xien Liu, Chenwei Yan, Xiao Zhang, Xinxin You, Yuxuan Zhou, Xiangling Fu, and Ji Wu. Critic–adviser–reviser cyclic refinement: Towards high-quality EMR corpus generation with LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7y11BdJIOp.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. De-identification is not enough: A comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 14(1):29669, 2024. doi: 10.1038/s41598-024-81170-y.

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.

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

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023. doi: 10.1038/s41746-023-00879-8.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15496–15523. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.916.

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, 2022. doi: 10.1038/s41467-022-35295-1.

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 O. Arik, and Tomas Pfister. EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine*, 6(1):141, 2023. doi: 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. 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, 2021. doi: 10.1093/jamia/ocaa262.

A Dataset Construction

We evaluate on two diagnosis-stratified clinical-note corpora. Table 3 summarizes the dataset statistics. Within each corpus, every method receives the same normalized age, sex, ordered text fields, and primary-diagnosis group, providing a common contract for generation and diagnosis-stratified evaluation.

Table 3: Dataset statistics used in all experiments. MIMIC-47 keeps all admissions from each patient in one partition.

Dataset	Groups	Records (train/test)	Patients (train/test)	Split rule	Note fields
EMR-302	302	153,842 / 38,308	–	record 80/20 within diagnosis	4
MIMIC-47	47	45,966 / 10,913	30,916 / 7,549	patient-disjoint 80/20	6

A.1 EMR-302

EMR-302 contains deidentified inpatient records from 302 primary-diagnosis groups. We use the established diagnosis-stratified 80/20 split, yielding 153,842 training and 38,308 test records. Each normalized record contains chief complaint, history of present illness, hospital course, and discharge instructions. Using the established split preserves direct comparability with prior work, while diagnosis-wise stratification ensures that every class evaluated by the downstream task is represented in both partitions.

A.2 MIMIC-47

MIMIC-47 is constructed from discharge summaries with one identifiable primary billed diagnosis. We extract six consistently available top-level sections: chief complaint, history of present illness, pertinent results, brief hospital course, discharge medications, and discharge instructions. Admissions are split at the patient level, with every admission from one patient kept in the same partition. This boundary prevents repeated admissions from the same individual from contributing to both guideline induction and final evaluation, yielding an admission-independent test population.

We exclude diagnosis blocks representing symptoms and signs, health-service contact factors, and complications of medical or surgical care so that each group denotes a clinically interpretable primary disease category. ICD-9-CM and ICD-10-CM codes are merged only through reciprocal, unique, exact, non-approximate, non-combination General Equivalence Mapping pairs. This conservative rule avoids label noise from ambiguous cross-version mappings and preserves a clear primary-diagnosis condition across all methods and evaluations.

Our metrics are estimated within diagnosis before macro-aggregation. A diagnosis represented by only a handful of records would therefore yield empirical entity distributions and diagnosis accuracy that are dominated by individual cases. To ensure that each diagnosis supports a stable diagnosis-stratified estimate, we fix the cohort eligibility rule before any model comparison and retain groups with at least 400 training patients and 100 test patients. The resulting 47 groups contain 45,966 training records from 30,916 patients and 10,913 test records from 7,549 patients. This rule defines the resolution at which the evaluation is reliable; all retained diagnoses subsequently receive equal weight in the macro metrics.

A.3 Training–Test Boundary

All guideline induction, source measurement, baseline guideline construction, and sampling-model fitting use only training records. The final synthetic corpus matches the held-out record count of every diagnosis group, and test note text is used only by the frozen evaluations. Our generator receives learned guidelines and sampled content plans rather than source-note text; only the two explicitly example-conditioned baselines receive training examples during generation. This common boundary isolates the value of learned guidelines while preserving a clean final comparison.

B Baseline Methods

The baselines represent typical ways of combining an LLM with fixed generation inputs: no additional guidelines, disease knowledge with a training example, topic and style guidelines with a training example, and iterative revision against predefined criteria. Within each dataset–backbone cell, all baselines use the same backbone, primary-diagnosis and demographic conditions, ordered output schema, and per-diagnosis corpus size. All resources supplied to the baselines are derived from training records without access to held-out test text, preserving the common training–test boundary.

B.1 Direct

Direct receives only the primary diagnosis, age, sex, and ordered field schema, and generates the record in one pass.

B.2 MedSyn

MedSyn receives a same-diagnosis training example together with symptom or finding concepts constructed from training records for that diagnosis as fixed generation guidelines.

B.3 ClinGen

ClinGen receives a same-diagnosis training example, a clinical topic constructed from training records, and a style description induced from training notes. Style induction excludes patient-specific facts and clinical content.

B.4 LLM-CARe

LLM-CARe retains its Critic–Adviser–Reviser organization at corpus, section, and document levels, with at most two revision cycles per level. It receives predefined clinical-quality criteria and training-derived aggregate targets but no real-record text during generation. Dataset field names are mapped to their corresponding semantic roles.

C Evaluation Metrics

C.1 Entity Frequency and Co-occurrence

The main-table evaluation uses a training-derived entity inventory shared by all methods and backbones within each dataset. For each diagnosis, an eligible entity must appear in at least 50 training records and be absent from at least 50; we retain up to 24 eligible entities in descending training-frequency order. Eligible pairs must co-occur in at least 20 training records, with at least 10 records containing only the first entity and 10 containing only the second. These entities and pairs are frozen before scoring any synthetic corpus and are not selected from learned guidelines or test results. The same text-matching procedure is applied to real and synthetic records, so both sides measure literal entity mentions using the same matching rules and record fields.

Let $\mathcal{Y}$ denote the set of diagnosis groups, $\mathcal{E}_g$ the frozen entity inventory for group $g \in \mathcal{Y}$, and $\mathcal{Q}_g$ its eligible unordered entity pairs. For $X \in \{\mathrm{real}, \mathrm{syn}\}$, let $\mathcal{D}_g^X$ denote the real test records or the synthetic records assigned to g, respectively, and let $\delta_{g,u}(e) \in \{0, 1\}$ indicate whether entity $u \in \mathcal{E}_g$

is mentioned in record e. Its empirical mention rate is $p^X_{g,u} = |\mathcal{D}^X_g|^{-1} \sum_{e \in \mathcal{D}^X_g} \delta_{g,u}(e)$. Entity frequency is evaluated by mean absolute error (MAE):

$$\mathrm{Freq} = \frac{1}{|\mathcal{Y}|} \sum_{g \in \mathcal{Y}} \frac{1}{|\mathcal{E}_g|} \sum_{u \in \mathcal{E}_g} \left| p^{\mathrm{real}}_{g,u} - p^{\mathrm{syn}}_{g,u} \right|. \tag{7}$$

For a pair $(u, v) \in \mathcal{Q}_g$, define $p^X_{g,uv}(\alpha, \beta)$ as the fraction of records in $\mathcal{D}^X_g$ with $\delta_{g,u}(e) = \alpha$ and $\delta_{g,v}(e) = \beta$, where $\alpha, \beta \in \{0, 1\}$. Entity co-occurrence uses total variation distance (TVD):

$$\mathrm{Cooc} = \frac{1}{|\mathcal{Y}|} \sum_{g \in \mathcal{Y}} \frac{1}{|\mathcal{Q}_g|} \sum_{(u,v) \in \mathcal{Q}_g} \frac{1}{2} \sum_{\alpha, \beta \in \{0,1\}} \left| p^{\mathrm{real}}_{g,uv}(\alpha, \beta) - p^{\mathrm{syn}}_{g,uv}(\alpha, \beta) \right|. \tag{8}$$

The joint TVD measures agreement across all four occurrence states and therefore reflects both marginal frequencies and co-occurrence. All 302 EMR-302 and 47 MIMIC-47 diagnosis groups have eligible entities and pairs. Evaluation includes every real test record and the same number of synthetic records per diagnosis: 38,308 records per method on EMR-302 and 10,913 on MIMIC-47.

C.2 Semantic Similarity and Coverage

We encode the complete canonical text fields in a fixed shared order with frozen Qwen3-Embedding-0.6B, without adding diagnosis labels or query instructions. Encoding uses BF16 inference and last-valid-token pooling; the resulting 1,024-dimensional vectors are stored in FP32 and normalized to unit length. The longest input contains 11,764 tokens, below the 32,768-token context limit, so all records are encoded without truncation or chunking. Both semantic measures use all real test records and all synthetic records, with identical diagnosis quotas across methods. Each measure is computed within diagnosis and then averaged equally over diagnoses.

For diagnosis g, let $n_g = |\mathcal{D}^{\mathrm{real}}_g|$ and $m_g = |\mathcal{D}^{\mathrm{syn}}_g|$ be the real and synthetic record counts, with $\mathcal{D}^X_g$ defined as in the entity evaluation, and let $\phi(e) \in \mathbb{R}^{1024}$ denote the unit embedding of record e. We use the linear kernel $\kappa(e, e') = \phi(e)^\top \phi(e')$ for records e, e'. With e, e' ranging over $\mathcal{D}^{\mathrm{real}}_g$ and $\tilde{e}, \tilde{e}'$ over $\mathcal{D}^{\mathrm{syn}}_g$, the unbiased squared maximum mean discrepancy is

$$\begin{aligned} \widehat{\mathrm{MMD}}^2_g = &\frac{\sum_{e \neq e'} \kappa(e, e')}{n_g(n_g - 1)} + \frac{\sum_{\tilde{e} \neq \tilde{e}'} \kappa(\tilde{e}, \tilde{e}')}{m_g(m_g - 1)} \\ &- \frac{2 \sum_e \sum_{\tilde{e}} \kappa(e, \tilde{e})}{n_g m_g}, \end{aligned} \tag{9}$$

With this kernel, MMD^2 measures the squared difference between mean semantic features; the unbiased estimate may be slightly negative for finite samples. It compares sets of records without requiring a one-to-one correspondence between real and synthetic examples.

Semantic coverage is measured by the real-to-synthetic nearest-neighbor cosine distance

$$\mathrm{Cov}_g = \frac{1}{n_g} \sum_{e \in \mathcal{D}^{\mathrm{real}}_g} \left[1 - \max_{\tilde{e} \in \mathcal{D}^{\mathrm{syn}}_g} \phi(e)^\top \phi(\tilde{e}) \right]. \tag{10}$$

Lower distance means real-record content is better represented by the synthetic set. This evaluates diversity through coverage of real cases, rather than dispersion among synthetic records alone. Keeping the synthetic count fixed within each diagnosis ensures that methods have equally sized nearest-neighbor candidate sets. The main table reports $|\mathcal{Y}|^{-1} \sum_{g \in \mathcal{Y}} \widehat{\mathrm{MMD}}^2_g$ and $|\mathcal{Y}|^{-1} \sum_{g \in \mathcal{Y}} \mathrm{Cov}_g$, using the same diagnosis groups as the entity evaluation.

C.3 Diagnosis Prediction

Diagnosis prediction uses train-synthetic-test-real transfer, a common evaluation of whether synthetic clinical records can replace real data for downstream training. We construct four-choice questions from the chief complaint and history of present illness, which contain the principal admission information while excluding diagnosis-revealing cues in later treatment and discharge fields. Each

question contains the primary diagnosis and three diagnosis-group distractors; distractors and option order are fixed and shared across methods, giving every method identical real test questions. We fine-tune a recent compact model, Qwen3.5-0.8B, on each synthetic corpus and test it on untouched real records. The probe and training recipe remain fixed across methods, so differences reflect the training corpus. Diagnosis-macro accuracy gives every diagnosis equal weight and is reported on $[0, 1]$ in the main table. Invalid answers count as incorrect.

C.4 Privacy Evaluation

The privacy analysis complements the main comparison by measuring record-reuse signals created by generation. Nearest-record similarity computes character 5-gram Jaccard similarity against up to 100 same-diagnosis training records and reports the 99th percentile across the synthetic corpus. The diagnosis-matched membership attack represents similarity to the synthetic corpus with character 3-, 4-, and 5-gram TF–IDF features. Member and nonmember records are paired within diagnosis using sex, age, and note length before a five-fold cross-fitted logistic classifier is trained, preventing obvious cohort imbalance from driving the attack. The reported membership score is the absolute ROC-AUC distance from chance.

D Implementation Details

D.1 Backbones

Qwen3.5-9B, Gemma-4-12B-it, and Llama-3.1-8B-Instruct each serve as the shared backbone for the Inducer and Generator in a separate experimental setting. We use each model's native chat template and provider-recommended sampling profile where available, otherwise retaining its standard model-family profile. Fixing these profiles before evaluation avoids both cross-family disadvantage and result-based tuning; methods sharing a backbone receive the same configuration. Qwen runs in non-thinking mode. Exact decoding settings will be provided with the released code.

D.2 Content-Guideline Induction

Content-guideline induction begins with an empty bank and repeatedly contrasts real records with records generated under the current bank. Refreshing the synthetic side removes addressed discrepancies and directs later proposals toward residual source structure. The Inducer proposes field-specific observable content, and each candidate is represented as a bounded regular-expression detector and replayed over all training records in its diagnosis. Separating LLM proposal from programmatic measurement prevents a plausible description from becoming evidence by itself. Requiring source recurrence also prevents a detail unique to one contrast from becoming a content guideline, while complete diagnosis-group replay supplies the measurements for planning.

For diagnosis g, the retained detectors form a binary training-record matrix. We estimate their pairwise mutual information and fit a diagnosis-specific Chow–Liu tree (Chow & Liu, 1968), a tree-structured approximation of their joint distribution that preserves the strongest pairwise dependencies and can be sampled efficiently. For every synthetic record, the planned number of content items is sampled from the empirical training distribution, after which the tree is sampled conditional on that exact cardinality. Preserving cardinality prevents guidance from becoming unnaturally dense, while joint sampling retains the strongest source dependencies. The draft is checked with the same detectors; a missing planned item is repaired within its assigned field only when the edit realizes that item and preserves every planned item already present in the draft.

D.3 Documentation-Guideline Induction

Documentation-guideline induction likewise repeats diagnosis-local contrasts after applying earlier guidelines, allowing repaired gaps to reveal finer residual requirements. Each proposal uses multiple real–synthetic pairs and separates an applicability condition C from a target T and editable fields. Entry screening requires a real $C{+}T{+}$ witness and a synthetic $C{+}T{-}$ witness, establishing source support and an actionable residual generation gap before broader validation.

Validation compares estimated target satisfaction rates on real training records excluded from proposal examples and on synthetic records from the current round. Within each diagnosis group, each rate is computed only over records satisfying the candidate condition, as in Equation 2. A candidate is admitted only when the real-record rate exceeds both the admission threshold $\tau = 0.8$ and the synthetic-record rate. This direct comparison retains requirements that recur in applicable real cases and remain less consistently satisfied by the current generator.

D.4 Induction Schedule

Both guideline types are induced within each diagnosis group under the same schedule. In every round, the Inducer contrasts four real training records with four synthetic records generated under the current bank. Induction stops once five consecutive rounds yield no newly validated guideline, or after at most 100 rounds, and the resulting bank is then frozen for final generation.