A Comparative Analysis of Large Language Models for Medical Tabular Synthetic Data Generation

Arshia Ilaty
San Diego State University
ailaty3088@sdsu.edu

Hossein Shirazi
San Diego State University
hshirazi@sdsu.edu

Hajar Homayouni
San Diego State University
hhomayouni@sdsu.edu

Abstract

Access to real-world medical data is often restricted due to privacy regulations, posing a significant barrier to the advancement of healthcare research. Synthetic data offers a promising alternative; however, generating realistic, clinically valid, and privacy-conscious records remains a major challenge. Recent advancements in Large Language Models (LLMs) offer new opportunities for structured data generation; however, existing approaches frequently lack systematic prompting strategies and comprehensive, multi-dimensional evaluation frameworks. In this paper, we present **SynLLM**, a modular framework for generating high-quality synthetic medical tabular data using 20 state-of-the-art open-source LLMs, including LLaMA, Mistral, and GPT variants, guided by structured prompts. We propose four distinct prompt types, ranging from example-driven to rule-based constraints, that encode schema, metadata, and domain knowledge to control generation without model fine-tuning. Our framework features a comprehensive evaluation pipeline that rigorously assesses generated data across statistical fidelity, clinical consistency, and privacy preservation. We evaluate SynLLM across three public medical datasets, including *Diabetes*, *Cirrhosis*, and *Stroke*, using 20 open-source LLMs. Our results show that prompt engineering significantly impacts data quality and privacy risk, with rule-based prompts achieving the best privacy-quality balance. SynLLM establishes that, when guided by well-designed prompts and evaluated with robust, multi-metric criteria, LLMs can generate synthetic medical data that is both clinically plausible and privacy-aware, paving the way for safer and more effective data sharing in healthcare research.

1 Introduction

Access to real-world medical data is frequently restricted due to privacy regulations, ethical constraints, and institutional barriers, posing a significant challenge for the development of AI-driven healthcare solutions. While data protection laws such as the Health Insurance Portability and Accountability Act (HIPAA) [11] and the General Data Protection Regulation (GDPR) [37] are essential for safeguarding patient confidentiality, they often hinder the availability of data for clinical model development and research. Synthetic data offers a promising alternative by enabling the training and validation of machine learning models without exposing real patient records.

Existing approaches to structured synthetic data generation, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and more recently, Large Language Models (LLMs), have shown potential but suffer from key limitations. GAN-based methods such as CT-GAN [41] and MedGAN [22] frequently experience mode collapse and require large amounts of real training data, limiting their utility in privacy-sensitive contexts [15]. VAEs tend to oversmooth feature distributions, thereby suppressing rare but clinically important conditions [13]. Addition-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

ally, both GANs and VAEs often struggle to capture complex feature interdependencies, resulting in synthetic records that lack medical plausibility.

Recent advancements in LLMs, including GReaT [8], and REaLTabFormer [31], present new opportunities for generating high-quality and privacy-preserving structured synthetic data. When guided with structured prompts, LLMs can produce contextually rich and statistically aligned tabular data. However, current LLM-based approaches face critical challenges:

Lack of structured prompting. Most existing methods rely on unstructured text generation followed by post-processing to construct tabular data, which is an additional overhead and can introduce errors.

Privacy risks. Without explicit and effective design constraints, LLMs may memorize and inadvertently replicate sensitive training records.

This work aims to investigate how prompt structure affects the quality and privacy of LLM-generated synthetic medical data. Specifically, we (1) develop a set of prompt strategies that encode schema information, statistical metadata, and clinical logic; (2) evaluate the ability of open-source LLMs to generate realistic and privacy-preserving synthetic records under these prompts; and (3) quantify performance trade-offs using a multidimensional evaluation framework that spans statistical fidelity, medical plausibility, and privacy risk.

To study how prompt structure affects synthetic data generation, we introduce **SynLLM**, a prompt-driven evaluation framework for structured medical data synthesis using LLMs. SynLLM implements four systematically designed prompt types, ranging from minimal information prompts that provide only column headers and a few example records to metadata-augmented and rule-based prompts that incorporate statistical summaries and domain-specific clinical constraints. Notably, the final prompt type excludes all example records and relies solely on rule-based guidance, allowing us to evaluate model performance under stricter privacy-aware generation conditions. These prompts guide LLMs in generating structured tabular records without requiring model fine-tuning. This design enables controlled comparisons of prompt effectiveness and supports the analysis of how different prompting strategies influence data quality, clinical validity, and privacy.

SynLLM is evaluated across three public medical datasets: *Diabetes*, *Cirrhosis*, and *Stroke* using 20 open-source LLMs, including Mistral-7B, Zephyr-7B, LLaMA, and GPT-2. Results demonstrate that prompt structure significantly impacts output quality and privacy. Rule-based prompts consistently achieve high harmonic privacy-quality scores without relying on example records. Our evaluation reveals that model behavior varies substantially across prompt types, highlighting the importance of prompt design in LLM-guided synthetic data generation.

Section 2 reviews relevant literature in synthetic data generation. Section 3 introduces the SynLLM pipeline and prompt types. Section 3.4 presents experimental setup and evaluation metrics. Section 4 provides empirical results and analysis. Section 5 provides key observations, followed by conclusions and future directions.

2 Related Work

Recent advancements in LLMs have demonstrated their ability to generate structured medical data by capturing complex feature interdependencies. GReaT introduced text-based encoding for tabular records, improving data diversity; however, with computational overhead and privacy risks. HARMONIC [40] presented instruction-tuned LLMs with k-nearest neighbors strategies that improved privacy preservation, though its evaluation metrics lack granularity in detecting structured privacy violations. Traditional models such as GANs (medGAN) and CTGAN improved categorical variable handling but suffer from mode collapse, computational intensity, and training sensitivity. VAEs provide smooth latent representations but generate overly averaged data, missing rare but critical cases. Diffusion models like TabDDPM [20] enhance distributional accuracy but require extensive computational resources.

Privacy-preserving techniques include DP-integrated methods, including DP-SDG [27], DP-GAN [18], and DP-WGAN [19] that inject noise into training procedures but often degrade synthetic data utility. Recent DP-enhanced LLM models like DP-LLMTabGen [36] show promise in balancing privacy and statistical fidelity. In contrast, our proposed SynLLM framework addresses these limitations through structured prompt engineering that embeds clinical logic and statistical properties explicitly. This approach maintains medical coherence, reduces computational overhead, and eliminates the need for latent-space modeling while enforcing metadata properties and domain-specific rules at generation time. SynLLM provides greater flexibility through prompt-based generation without requiring model retraining for different subpopulations.

3 Methodology

SynLLM is built around a modular pipeline that includes schema profiling, prompt construction, LLM-based record generation, and multi-dimensional evaluation. The core methodological innovation lies in the use of structured, domain-informed prompts that guide generation without requiring model retraining or fine-tuning. We describe the four prompt strategies employed, the data generation process across 20 LLMs, and the multi-dimensional evaluation criteria used to assess statistical fidelity, clinical consistency, privacy risk, and computational efficiency.

3.1 Problem Definition

Let $\mathcal{D}_{\text{real}} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote a structured electronic health record (EHR) dataset, where each row $x^{(i)} \in \mathbb{R}^{p_{\text{num}}} \times \mathcal{C}^{p_{\text{cat}}}$ comprises p_{num} numerical and p_{cat} categorical attributes, and $y^{(i)}$ is an optional downstream label. We define a prompt-driven generation mechanism $\mathcal{G}_\theta : (\Pi,\, k) \longmapsto \hat{\mathcal{D}}_{\text{syn}}$ that, given a prompt specification Π and target record count $k \ll N$, produces a synthetic dataset $\hat{\mathcal{D}}_{\text{syn}}$ such that:

Statistical fidelity: $\hat{\mathcal{D}}_{\text{syn}}$ approximates the marginal and joint distributions of $\mathcal{D}_{\text{real}}$ within a tolerance $\varepsilon_{\text{stat}}$.

Clinical plausibility: Synthetic records satisfy logical and medical constraints (e.g., HbA1c $>$ 6.5 $\Rightarrow$ Diabetes = True).

Privacy preservation: The probability that any $\hat{x} \in \hat{\mathcal{D}}_{\text{syn}}$ is linkable to a real record is bounded above by δ_{priv}, as estimated via empirical privacy metrics (e.g., k-anonymity, membership inference, nearest-neighbor distance).

Unlike GAN- or VAE-based methods, which require access to real patient records during model training, SynLLM leverages zero- and few-shot LLM inference guided by carefully designed prompts. These prompts incorporate only aggregate statistics and domain rules extracted from $\mathcal{D}_{\text{real}}$, without exposing any individual-level data. By using only non-identifiable summaries—feature distributions, clinical thresholds, and correlations—SynLLM minimizes disclosure risk while leveraging the rich priors of instruction-tuned language models [8, 31].

3.2 SynLLM Framework Overview

The SynLLM pipeline (Algorithm 1) consists of four modular stages that enable LLM-based generation of privacy-conscious, clinically meaningful structured medical data. Each stage is designed to preserve fidelity to real data characteristics while minimizing privacy risks.

Schema Analysis. Extract attribute types, univariate statistics (e.g., mean, standard deviation, min–max range), and relevant inter-feature correlations from $\mathcal{D}_{\text{real}}$ (Sec. 3.3). Only aggregated metadata, never raw records, are surfaced outside the secure data enclave.

Prompt Construction. Construct a generation prompt Π using one of four progressively constrained templates, each encoding different levels of statistical metadata and clinical logic (Sec. 3.3).

LLM Inference. Query an instruction-tuned, open-source language model (see Table 6) using fixed sampling parameters (temperature T=0.7, top-p=0.9). The token budget is dynamically adjusted based on the desired record count k.

Post-processing and Validation. Parse generated JSON objects into structured tabular form, enforce data typing constraints, and discard records violating hard-coded clinical rules. Validated records are passed to the evaluation pipeline described in Sec. 4.

```
**Input:** Real dataset $\mathcal{D}_{\text{real}}$ (for schema extraction only), set of LLMs $\mathcal{M}$, prompt templates $\mathcal{P}$
**Output:** Synthetic dataset $\hat{\mathcal{D}}_{\text{syn}}$ with statistical, clinical, and privacy evaluations
**Stage 1: Metadata Extraction**
Extract feature schema $\mathcal{S}$, value ranges, types, and statistical summaries from $\mathcal{D}_{\text{real}}$;
Identify domain rules and clinical constraints $\mathcal{R}$ from medical knowledge base or expert guidance;
**Stage 2: Prompt Engineering**
**foreach** *prompt type* $p \in \mathcal{P}$ **do**
    Construct prompt $P$ using schema $\mathcal{S}$, metadata, and rules $\mathcal{R}$;
**end**
**Stage 3: Synthetic Data Generation**
**foreach** *model* $m \in \mathcal{M}$ **do**
    **foreach** *prompt* $P$ **do**
        Generate synthetic records $R_{m,P} = m(P)$;
        Parse $R_{m,P}$ into structured tabular format;
    **end**
**end**
**Stage 4: Evaluation and Filtering**
**foreach** *synthetic record set* $R_{m,P}$ **do**
    Compute statistical metrics (e.g., Wasserstein, correlation);
    Compute medical consistency scores based on $\mathcal{R}$;
    Compute privacy risk metrics (e.g., k-anonymity, NN distance);
    Optionally filter or flag low-quality or high-risk records;
**end**
**return** $\hat{\mathcal{D}}_{syn} = \bigcup R_{m,P}$
```

Algorithm 1: SynLLM: Structured Medical Data Generation with LLMs

3.3 Adaptive Prompt Taxonomy

Our prompt schema is organized into a four-tier hierarchy of escalating sophistication: Level 1 functions as the baseline, while levels 2 through 4 incrementally introduce richer contextual cues, including feature definition and statistical properties, and stricter domain-specific constraints.

SeedEx (Prompt-A): *Example-Seed Minimal Prompt.* Lists the column headers corresponding to dataset features, the desired output format, and ≤ 5 seed rows randomly sampled from $\mathcal{D}_{\text{real}}$, not generated or anonymized. Purpose: to establish a *baseline* that stresses model generalization under minimal constraint. However, this formulation presents the highest risk of record memorization and identity disclosure.

FeatDesc (Prompt-B): *Feature-Description Prompt.* Replaces concrete examples with concise natural-language definitions of each attribute (e.g. "`bmi`: body-mass index in kg/m^2, a continuous variable bounded within $[12, 60]$). This approach introduces semantic structure by providing the model with descriptive, clinically grounded definitions of each feature, which guide the generation process and help constrain outputs to realistic, in-distribution value ranges.

StatGuide (Prompt-C): *Statistical-Metadata Prompt.* Extends FeatDesc with feature-level summaries including means, standard deviations, min–max bounds, category frequencies, and selected pairwise correlations. This template draws inspiration from the "data portrait" concept in [40], which encodes statistical summaries to guide generation. In our framework, we apply similar dataset-level metadata to construct the StatGuide prompt, which has been empirically shown to reduce divergence from the target distribution (Sec. 4).

ClinRule (Prompt–D): *Clinically-Constrained Prompt.* Eliminates example records entirely and replaces them with declarative logic rules derived from medical guidelines (e.g., "If `pregnant`=True, then `sex`=Female"). The LLM is required to generate samples that satisfy these constraints, thereby prioritizing logical consistency and minimizing disclosure risk.

Table 1 provides an abridged overview of each prompt template. All prompts share a consistent **system message** instructing the model to (i) emit `newline`-delimited JSON objects, (ii) avoid free-text commentary, (iii) adhere to the requested number of records, and (iv) refrain from emitting any protected health information (*PHI*). During prompt construction, dataset-specific metadata is programmatically inserted into placeholder tags (e.g., `{feature_stats}`).

Table 1: Prompt skeletons (abridged). Curly braces denote runtime placeholders.

Template	Key Sections
SEEDEX	Header row; n example records; “Repeat format exactly, k rows.”
FEATDESC	Header row; per-feature descriptions; JSON schema block.
STATGUIDE	As FEATDESC, plus {mean}, {stdev}, {min,max}, frequency tables; optional correlation matrix snippet.
CLINRULE	Header row; domain-specific logic rules (e.g., DL $\rightarrow$ HbA1c $>$ 6.5); JSON schema; no examples.

Design Rationale This prompt taxonomy systematically varies the amount and type of conditioning information supplied to the LLM, allowing for controlled exploration of the privacy–utility trade-off. SEEDEX provides minimal constraint, often resulting in low Jensen–Shannon divergence but elevated membership inference risk. At the opposite end, CLINRULE imposes strict domain rules, substantially mitigating privacy risk at the expense of greater distributional shift. The intermediate templates, FEATDESC and STATGUIDE, introduce semantic and statistical context, enabling precise evaluation of how information content affects fidelity and generalization. Empirical results in Sec. 4 show that STATGUIDE achieves the best utility for internal analytics, while CLINRULE is most suitable for public release scenarios.

Schema and Statistical Extraction For each numerical attribute f, we extract the 5-tuple $(\mu_f, \sigma_f, \min_f, \max_f, \text{quantiles}_f)$. For categorical attributes, we compute the empirical probability mass function $\mathbf{p}_f$. To reduce the risk of rare-category disclosure, we apply a frequency threshold of five and consolidate infrequent values into an “`Other`” category before incorporating them into prompt metadata. Pairwise Pearson correlations ρ_{fg} are retained only if $|\rho_{fg}| > 0.15$ or identified as clinically relevant by domain experts.

3.4 Evaluation and Metrics Description

To evaluate the effectiveness of SynLLM, we conduct a comprehensive **quality–privacy–utility audit** that assesses each synthetic dataset across four orthogonal performance dimensions: **Statistical fidelity**, **Clinical consistency**, **Privacy protection**, and **Machine learning utility**.

3.4.1 Statistical Fidelity Evaluation

We evaluate statistical fidelity using distributional and relational metrics across three dimensions:

Marginal Distribution Alignment: Wasserstein distance [38], Jensen–Shannon divergence [23], Anderson–Darling k-sample test [30], Kullback–Leibler divergence [21], and range coverage.

Dependency Preservation: Pearson correlation coefficients [25], Frobenius norm of correlation matrix differences [14], and feature-level correlation analysis [42] for medically relevant relationships.

Categorical Structure: χ^2 test [26], category preservation rate [10], and mutual information score [12] for co-dependence among categorical variables.

These metrics quantify fidelity from complementary angles: feature distribution realism, statistical dependency preservation, and categorical structure integrity.

3.4.2 Clinical Consistency Evaluation

We evaluate clinical consistency using domain-informed metrics based on epidemiological principles to ensure synthetic data preserves medically meaningful relationships.

Examples: HbA1c differences between diabetic/non-diabetic groups, glucose levels by stroke outcome, age-stroke risk gradients, and hypertension-stroke co-occurrence patterns. Deviations are computed using group-wise mean differences or regression slope variations relative to real data.

Scoring: Aggregated into a clinical consistency score where lower values indicate better alignment with expected clinical patterns, helping identify medically implausible outputs.

3.4.3 Privacy Protection Evaluation

SynLLM evaluates privacy via empirical, distance-based metrics that estimate resemblance between synthetic and real records—without formal guarantees.

Nearest Neighbor Distance Ratio. For each synthetic record, we compute its distance to the nearest real record and compare it to average real–real distances. A higher ratio indicates better privacy through greater separation.

Identifiability Score. Measures the fraction of synthetic records exactly matching real ones across all features. Lower is better, indicating less memorization.

These metrics offer interpretable, model-agnostic privacy signals. However, SynLLM does not implement formal protections like differential privacy or k-anonymity. Results should thus be viewed as an empirical audit rather than a privacy guarantee.

Synthetic data batches that exhibit high privacy risk scores or violate anonymity thresholds are logged for further analysis and may inform prompt refinement or post-processing strategies in subsequent iterations of the generation pipeline.

3.4.4 Machine Learning Utility Evaluation

Beyond distributional and clinical alignment, synthetic data must support downstream predictive tasks. We assess **machine learning utility** by testing whether models trained on synthetic data generalize comparably to real-data-trained counterparts, as detailed in Sec. 3.4.

We use three tree-based classifiers widely applied in healthcare due to their robustness and interpretability: a **Decision Tree** (depth 5), a **Random Forest** (50 trees), and an **XGBoost** model (100-round early stopping, default settings).

Generalization is evaluated via two complementary setups: **TSTR** (Train on Synthetic, Test on Real) and **TRTS** (Train on Real, Test on Synthetic), capturing forward and backward fidelity.

Primary metrics include accuracy, macro F1, and AUC-ROC. We supplement this with diagnostics such as confusion matrices, PR curves, and feature importance to assess where synthetic data helps or harms task performance.

4 Results and Analysis

Datasets: To evaluate the effectiveness of SynLLM in generating high-quality synthetic medical data, we conducted experiments on three publicly available, structured healthcare datasets. These datasets span distinct clinical domains, diabetes diagnosis, cirrhosis severity classification, and stroke prediction—and are commonly used in medical machine learning research. All are designed for binary or multi-class classification tasks. Detailed dataset statistics, including record counts, feature types, and target labels, are provided in Supplementary Table 5.

LLM Selection: We evaluated SynLLM across 20 open-source LLMs spanning diverse architectures, parameter sizes, and fine-tuning strategies (Table 6).

4.1 Focused Model Analysis: Privacy–Quality Trade-Off Across Prompt Variants

A central challenge in synthetic medical data generation is achieving a favorable balance between output quality and privacy protection. In SynLLM, we assess this trade-off by evaluating 20 LLMs under four distinct prompting strategies across three medical datasets. Table 2 reports normalized scores for quality, privacy, and their harmonic mean, serving as a composite indicator of overall generation efficacy.

Score Aggregation. We compute composite quality and privacy scores by averaging normalized metrics. Quality metrics (Wasserstein distance, correlation preservation) are directionally aligned so higher values reflect better fidelity: Quality $= \frac{1}{N}\sum_{i=1}^{N} \text{NormalizedQuality}_i$. Privacy metrics (nearest-neighbor ratios, identifiability scores) are similarly normalized to $[0, 1]$ with higher values indicating stronger protection: Privacy $= \frac{1}{M}\sum_{j=1}^{M} \text{NormalizedPrivacy}_j$. These composite scores enable unified model comparison in Section 4.1.2.

4.1.1 Prompt-Level Analysis

While SynLLM was evaluated on a broad set of 20 open-source LLMs, we present a focused analysis on five representative models: Zephyr 7B, OpenChat 7B, LLaMA 8B, Nous Hermes 34B, and GPT-2 variants. This subset was selected based on the following criteria:

Architectural diversity: The models span multiple LLM families (Zephyr, OpenChat, LLaMA, Yi, GPT) and include both recent instruction and chat-tuned architectures and established baselines. **Scale and alignment variation:** The selection includes small-scale (<1B), medium-scale (7–8B), and large-scale (34B) models with differing context lengths.

SEEDEX – Example-Based Prompting: *Diabetes:* Zephyr 7B leads in quality, while GPT-2-Large shows the highest privacy score but at a cost to fidelity. Most models display strong quality with moderate privacy, reinforcing that direct examples increase realism but elevate leakage risk. *Stroke:* OpenChat 7B performs best overall, achieving the highest quality. GPT-2-Large lags in both dimensions, while LLaMA 8B performs well on privacy but shows mixed quality outcomes. *Cirrhosis:* OpenChat 7B again tops quality, while Zephyr 7B leads in privacy. LLaMA 8B and Nous Hermes trail in privacy but maintain high quality.

FEATDESC – Feature Definition Prompt: *Diabetes:* Zephyr and Nous Hermes show the best balance. LLaMA 8B retains relatively high privacy but shows weaker quality. The shift from examples to definitions improves privacy for most models with minor loss in fidelity. *Stroke:* LLaMA 3.1 8B achieves the highest privacy performance, while Nous Hermes Yi 34B leads in quality. OpenChat 7B offers a strong balance between quality and privacy. In contrast, GPT-2 variants perform the worst. *Cirrhosis:* OpenChat 7B achieves near-perfect quality; however, Zephyr 7B provides the balance between privacy and quality. GPT-2 results remain the worst.

STATGUIDE – Metadata-Augmented Prompt: *Diabetes:* Quality is more consistent across models, with Zephyr, OpenChat, and Nous Hermes performing similarly. GPT-2-Large achieves top privacy but lower quality, highlighting trade-off extremes. *Stroke:* OpenChat and Nous Hermes achieve the highest quality scores, while also maintaining reasonably consistent and acceptable levels of privacy. In contrast, GPT-2 continues to exhibit poor fidelity, failing to generate outputs aligned with clinical expectations. These findings suggest that structured metadata guidance is sufficient to enhance quality without compromising privacy. *Cirrhosis:* Zephyr leads in both quality and privacy; OpenChat follows closely.

CLINRULE – Rule-Based Prompting: *Diabetes:* Zephyr, OpenChat, and Nous Hermes exhibit consistently strong performance in terms of quality. While privacy scores remain relatively stable across these models, they tend to be modest in magnitude. In contrast, GPT-2 variants fail to generate valid outputs, likely due to their limited capacity and architecture. *Stroke:* OpenChat again excels, with Nous Hermes closely matched. GPT-2 remains unsupported under this prompt. *Cirrhosis:* OpenChat variants achieve the highest quality scores but exhibit the lowest privacy scores, highlighting a pronounced trade-off between fidelity and confidentiality. Most other models follow a similar pattern, with marginal differences.

4.1.2 Prompt Variation and Harmonic Score Trends

We compute a harmonic score to evaluate joint quality-privacy performance: $\text{Harmonic Score} = \frac{2QP}{Q+P}$, where Q and P are normalized quality and privacy scores. This metric emphasizes balanced performance by penalizing model–prompt pairs where one dimension significantly underperforms.

CLINRULE Outperforms in Privacy-Conscious Generation. CLINRULE consistently yields high harmonic scores across top-tier models. This result is especially significant because CLINRULE includes no real data examples—only domain rules and metadata—suggesting that well-designed, constraint-based prompting can deliver high-quality outputs with minimal privacy risk.

STATGUIDE Maximizes Quality but Sacrifices Privacy in Some Models. STATGUIDE leads to some of the highest individual quality scores as seen in the previous subsection.

SEEDEX and FEATDESC Show Model-Specific Sensitivity. While SEEDEX offers moderate performance for many models. FEATDESC provides a more consistent profile, improving performance for several models like OpenChat and Nous Hermes in stroke and cirrhosis datasets, but still falls short for foundational models (GPT-2 variants).

Table 2: Normalised scores for 20 LLMs under four prompting strategies across three medical datasets (Diabetes, Stroke, and Cirrhosis). Each prompt is evaluated on three metrics: *Quality*, *Privacy*, and their *harmonic*. Higher values are better.

Dataset	LLM	SeedEx			FeatDesc			StatGuide			ClinRule		
		Qual.	**Priv.**	**H-Avg.**	**Qual.**	**Priv.**	**H-Avg.**	**Qual.**	**Priv.**	**H-Avg.**	**Qual.**	**Priv.**	**H-Avg.**
Diabetes	Zephyr 7B	**0.77**	0.42	**0.59**	0.66	0.42	0.54	0.66	0.46	0.56	0.63	0.41	0.52
	OpenChat 3.5 GPTQ	0.63	0.42	0.52	0.64	0.42	0.53	0.67	0.37	0.52	0.63	0.53	0.58
	Nous Hermes Yi 34B	0.64	0.32	0.48	0.65	0.42	0.53	0.56	0.41	0.48	0.58	0.41	0.50
	OpenChat 3.5	0.68	0.40	0.54	0.65	0.38	0.52	0.66	0.43	0.55	0.64	0.38	0.51
	GPT-2-Large	0.63	**0.53**	0.58	0.39	0.32	0.36	0.51	**0.66**	**0.59**	–	–	–
	GPT-2-Medium	0.50	0.26	0.38	0.63	**0.52**	**0.57**	0.64	0.41	0.52	–	–	–
	GPT-2-Small	0.43	0.36	0.39	0.49	0.30	0.40	0.37	0.43	0.40	–	–	–
	Mistral 7B	0.51	0.38	0.45	0.58	0.40	0.49	0.55	0.44	0.49	0.64	0.57	0.60
	Qwen2 7B	0.62	0.37	0.50	0.61	0.27	0.44	0.55	0.21	0.38	0.60	0.44	0.52
	InternLM2.5 7B	0.61	0.39	0.50	0.63	0.35	0.49	0.55	0.21	0.38	0.62	0.54	0.58
	Yi 6B	0.55	0.27	0.41	0.63	0.37	0.50	0.43	0.29	0.36	0.53	**0.78**	0.65
	LLaMA 2 13B	0.68	0.31	0.49	**0.66**	0.33	0.50	0.69	0.33	0.51	**0.67**	0.26	0.46
	LLaMA 2 13B Chat	0.60	0.24	0.42	0.60	0.25	0.43	0.56	0.22	0.39	0.56	0.40	0.48
	LLaMA 3.1 8B	0.55	0.36	0.45	0.62	0.35	0.49	0.62	0.24	0.43	0.53	0.47	0.50
	Mosaic MPT 7B	0.57	0.21	0.39	0.54	0.23	0.39	0.58	0.24	0.41	0.62	0.71	**0.67**
	Gemma 7B	0.56	0.26	0.41	0.60	0.22	0.41	0.62	0.26	0.44	0.60	0.36	0.48
	Nous Hermes Mistral 7B	0.64	0.49	0.56	0.66	0.45	0.56	**0.71**	0.41	0.56	0.54	0.54	0.54
Stroke	Zephyr 7B	0.56	0.54	0.55	0.69	0.39	0.54	0.79	0.57	0.68	0.61	0.49	0.55
	OpenChat 3.5 GPTQ	0.71	0.54	0.62	0.78	0.57	0.67	0.80	0.52	0.66	0.83	0.44	0.63
	Nous Hermes Yi 34B	0.67	0.54	0.61	**0.88**	0.47	0.67	**0.87**	0.42	0.65	0.74	0.49	0.61
	OpenChat 3.5	**0.82**	0.52	0.67	0.77	0.67	**0.72**	0.83	0.60	0.71	**0.87**	0.56	**0.71**
	GPT-2-Large	0.54	0.32	0.43	0.51	0.30	0.41	0.20	0.40	0.30	–	–	–
	GPT-2-Medium	0.42	0.25	0.33	0.42	0.25	0.33	0.44	0.48	0.46	–	–	–
	GPT-2-Small	0.48	0.25	0.37	0.42	0.25	0.33	0.21	0.46	0.33	–	–	–
	Mistral 7B	0.70	0.51	0.60	0.60	0.53	0.57	0.81	0.65	**0.73**	0.87	0.43	0.65
	Qwen2 7B	0.59	0.41	0.50	0.51	0.46	0.49	0.53	0.40	0.46	0.42	**0.75**	0.58
	InternLM2.5 7B	0.66	0.40	0.53	0.74	0.58	0.66	0.59	**0.68**	0.63	0.51	0.48	0.50
	Yi 6B	0.75	**0.73**	**0.74**	0.80	0.52	0.66	0.60	0.43	0.52	0.65	0.71	0.68
	LLaMA 2 13B	0.43	0.26	0.35	0.42	0.25	0.33	0.62	0.50	0.56	0.41	0.33	0.37
	LLaMA 2 13B Chat	0.43	0.37	0.40	0.50	0.32	0.41	0.62	0.43	0.53	0.64	0.73	0.69
	LLaMA 3.1 8B	0.43	0.62	0.52	0.56	**0.69**	0.62	0.57	0.54	0.55	0.69	0.53	0.61
	Gemma 7B	0.60	0.30	0.45	0.69	0.54	0.61	0.28	0.30	0.29	0.55	0.55	0.55
	Nous Hermes Mistral 7B	0.65	0.51	0.58	0.56	0.51	0.53	0.76	0.50	0.63	0.64	0.52	0.58
Cirrhosis	Zephyr 7B	0.59	**0.75**	**0.67**	0.66	**0.68**	**0.67**	**0.86**	**0.39**	**0.63**	0.50	0.39	0.44
	OpenChat 3.5 GPTQ	0.80	0.44	0.62	0.82	0.39	0.60	0.61	0.34	0.47	0.88	0.26	0.57
	Nous Hermes Yi 34B	0.84	0.30	0.57	0.85	0.35	0.60	0.64	0.32	0.48	0.66	0.27	0.47
	OpenChat 3.5	**0.91**	0.42	0.67	**0.98**	0.34	0.66	0.72	0.34	0.53	**1.00**	0.26	0.63
	GPT-2-Small	0.14	0.25	0.20	0.00	0.25	0.12	0.00	0.25	0.12	–	–	–
	Qwen2 7B	0.65	0.43	0.54	0.76	0.35	0.55	0.42	0.28	0.35	0.74	0.28	0.51
	InternLM2.5 7B	0.68	0.50	0.59	0.70	0.41	0.55	0.52	0.29	0.40	–	–	–
	Yi 6B	0.22	0.28	0.25	0.41	0.39	0.40	0.50	0.31	0.41	–	–	–
	LLaMA 3.1 8B	0.81	0.36	0.59	0.79	0.30	0.54	0.61	0.36	0.49	0.75	**0.52**	**0.63**
	StableBeluga 7B	0.00	0.25	0.12	0.00	0.25	0.12	0.00	0.25	0.13	–	–	–
	Gemma 7B	0.55	0.31	0.43	0.68	0.29	0.49	0.00	0.25	0.13	0.94	0.26	0.60

4.2 Machine Learning Utility

We evaluate utility using two standard strategies (Sec. 3.4): Train-on-Synthetic, Test-on-Real (TSTR) and Train-on-Real, Test-on-Synthetic (TRTS), assessing whether synthetic data captures predictive structure and supports downstream modeling. Table 4 reports mean accuracy, macro F1, and AUC-ROC for the Diabetes dataset across all prompts.

Model performance varies with fidelity and privacy. Nous Hermes Yi 34B achieves high TSTR scores (AUC and accuracy > 0.91), while Yi 6B leads in TRTS AUC (≥ 0.98), indicating strong semantic alignment with real data. Zephyr 7B and OpenChat 7B show balanced performance in both settings (AUC ≥ 0.89). GPT-2 models perform well on TSTR and privacy but show weaker F1, likely due to limited capacity.

Overall, SynLLM's outputs preserve sufficient signal for reliable prediction, even without retraining or fine-tuning. Strong TSTR/TRTS scores confirm the framework's effectiveness in generating both high-utility and privacy-conscious synthetic records.

4.3 Model Efficiency Analysis

We introduce the **Global Fidelity Index (GFI)**, aggregating statistical fidelity, privacy, and medical consistency metrics. Model efficiency combines normalized generation speed and GFI: $\text{Efficiency Score} = \frac{1}{2}(\text{NormSpeed} + \text{GFI})$, where higher scores indicate better speed-quality trade-offs. Table 3 shows Nous Hermes 34B achieving optimal efficiency (0.078), followed by Zephyr 7B (0.093). OpenChat 7B and LLaMA 8B show strong fidelity but slower generation, while GPT-2

Table 3: Model Efficiency Ranking

Rank	**Model**	**Dur (s)**	**GFI**	**Eff.**
1	Nous Hermes 34B	133	.096	.078
2	Zephyr 7B	122	.101	.093
3	OpenChat 7B	1522	.098	.215
4	LLaMA 8B	2293	.091	.264
5	GPT-2	287	.166	.294

Table 4: Diabetes ML Utility (TSTR / TRTS)

Model	**TSTR**			**TRTS**		
	Acc	F1	AUC	Acc	F1	AUC
GPT-2-L	.90	.50	.83	.86	.81	.90
GPT-2-M	.92	.57	.85	.88	.76	.98
GPT-2-S	.90	.53	.86	.93	.87	**.99**
Gemma 7B	.90	.59	.87	.90	.89	.94
InternLM2.5 7B	.89	.53	.89	.89	.84	.98
LLaMA 3.1 8B	.92	.67	.91	.90	.84	.98
Mistral 7B	.90	.60	.88	.82	.77	.92
Nous Hermes 34B	**.93**	**.74**	**.92**	.87	.75	.94
OpenChat 3.5	.92	.70	.89	.86	.71	.85
Yi 6B	0.82	0.46	0.79	**0.98**	**0.96**	0.98
Zephyr 7B	.88	.54	.82	.88	.74	.89

ranks lower due to limited fidelity.**Note.** Full prompt-wise metrics will be available upon acceptance for reproducibility.

5 Key Observations and Discussion

Our evaluation across three datasets, four prompt strategies, and 20 open-source LLMs reveals that models such as OpenChat 7B, Zephyr 7B, and Nous Hermes 34B consistently rank among the top performers across statistical, clinical, and privacy metrics. Notably, the CLINRULE prompt, designed without any data examples, achieves the highest harmonic privacy–utility scores, demonstrating the effectiveness of constraint-driven generation under strong privacy requirements.

Structured Prompting as a Privacy–Utility Lever. A central finding is that prompt structure exerts significant influence on both data fidelity and privacy risk. Prompts using real data examples yield high TSTR and distributional scores; however, at the cost of increased privacy risk. In contrast, CLINRULE, which encodes only declarative clinical rules, preserves utility while drastically reducing memorization behavior. This supports SynLLM's design hypothesis that structured, constraint-aware prompting enables high-fidelity generation without reliance on direct example exposure.

Prompt Sensitivity and Model Robustness. Performance varies notably across models: instruction-tuned ones (e.g., OpenChat 7B, Zephyr-7B) handle diverse prompts well, whereas GPT-2 variants degrade under stricter constraints. This highlights the need for adaptive or automated prompt strategies tailored to model alignment and response patterns.

Multidimensional Evaluation and Limitations. SynLLM employs a comprehensive evaluation suite integrating univariate and multivariate statistical tests (e.g., Wasserstein distance, Frobenius norm), clinical plausibility checks, and empirical privacy audits (e.g., nearest-neighbor distance ratios, identifiability scores). While this framework enables rigorous model comparison, the privacy metrics remain heuristic and empirically grounded. Future work may incorporate formal differential privacy analysis or white-box adversarial testing to strengthen guarantees.

6 Conclusion and Future Directions

In this paper, we presented SynLLM, a privacy-aware framework for synthetic structured medical data generation using large language models. By leveraging dataset-derived metadata and declarative domain knowledge, SynLLM crafts structured prompts that guide LLMs in producing high-fidelity, clinically plausible, and privacy-preserving tabular records without requiring access to real patient data during inference. Our evaluation spans 20 open-source LLMs and four systematically designed prompt strategies across three public datasets, assessing statistical fidelity, clinical consistency, machine learning utility, and empirical privacy risk. The results confirm that prompt-only control can match or exceed the quality of GAN and VAE baselines, while drastically simplifying deployment and model reuse. Future improvements to SynLLM could explore adaptive prompt optimization strategies, including metric-guided or reinforcement learning-based prompt tuning. Expanding support for multimodal EHRs (e.g., clinical text, imaging) and investigating synergies with federated learning may further enhance privacy and utility. These directions will continue to strengthen SynLLM as a foundational tool for scalable and responsible synthetic data generation.

Acknowledgment: This research project has benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

References

[1] Cirrhosis Prediction Dataset, 2021.

[2] Stroke Prediction Dataset, 2021.

[3] Diabetes Prediction Dataset, 2023.

[4] 01.AI. Yi-6b chat. `https://huggingface.co/01-ai/Yi-6B-Chat`, 2023.

[5] Meta AI. Llama 3.1 8b instruct. `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`, 2024.

[6] Mistral AI. Mistral 7b instruct v0.2. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`, 2023.

[7] Stability AI. Stablebeluga 7b. `https://huggingface.co/stabilityai/StableBeluga-7B`, 2023.

[8] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators, 2023.

[9] Zheng Cai, Maosong Cao, Haojiong Chen, and et al. Internlm2 technical report, 2024.

[10] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *Machine Learning for Healthcare Conference*, pages 286–305, 2017.

[11] U.S. Congress. Health insurance portability and accountability act of 1996 (hipaa). `https://www.hhs.gov/hipaa/index.html`, 1996. Accessed May 2025.

[12] Thomas M Cover. Elements of information theory. *John Wiley & Sons*, 1999.

[13] Sanket Dash, Oktay Günlük, and Dennis Wei. Privacy-preserving synthetic medical data generation using variational autoencoders. *arXiv preprint arXiv:2012.15328*, 2020.

[14] Gene H Golub and Charles F Van Loan. Matrix computations. *Johns Hopkins Studies in Mathematical Sciences*, 2013.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[16] Google. Gemma 7b it. `https://huggingface.co/google/gemma-7b-it`, 2024.

[17] Hugging Face H4. Zephyr 7b beta. `https://huggingface.co/HuggingFaceH4/zephyr-7b-beta`, 2023.

[18] Stella Ho, Youyang Qu, Bruce Gu, Longxiang Gao, Jianxin Li, and Yong Xiang. Dp-gan: Differentially private consecutive data publishing using generative adversarial nets. *Journal of Network and Computer Applications*, 185:103066, 2021.

[19] Jiaqi Huang, Qiushi Huang, Gaoyang Mou, and Chenye Wu. Dpwgan: High-quality load profiles synthesis with differential privacy guarantees. *IEEE Transactions on Smart Grid*, 14(4):3283–3295, 2023.

[20] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[21] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[22] Bruno Macedo, Inês Ribeiro Vaz, and Tiago Taveira Gomes. Medgan: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Scientific Reports*, 14(1):1212, 2024.

[23] M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[24] MosaicML. Mpt-7b-instruct. `https://huggingface.co/mosaicml/mpt-7b-instruct`, 2023.

[25] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[26] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[27] Le Trieu Phong and Tran Thi Phuong. Differentially private stochastic gradient descent via compression and memorization. *Journal of Systems Architecture*, 135:102819, 2023.

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[29] Nous Research. Nous-hermes-2-yi-34b. `https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B`, 2024.

[30] Fritz W Scholz and Michael A Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[31] Aivin V Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*, 2023.

[32] OpenChat Team. Openchat 3.5. `https://huggingface.co/openchat/openchat_3.5`, 2024.

[33] Qwen Team. Qwen-7b chat. `https://huggingface.co/Qwen/Qwen1.5-7B-Chat`, 2023.

[34] Qwen Team. Qwen2-7b instruct. `https://huggingface.co/Qwen/Qwen2-7B-Instruct`, 2024.

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[36] Toan Tran and Li Xiong. Differentially private tabular data synthesis using large language models. 06 2024.

[37] European Union. Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). `https://gdpr-info.eu`, 2016. Accessed May 2025.

[38] Cédric Villani. *Optimal transport: old and new*. Springer, 2009.

[39] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

[40] Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. Harmonic: Harnessing llms for tabular data synthesis and privacy protection. *ArXiv*, abs/2408.02927, 2024.

[41] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[42] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Supplementary

This appendix aims to foster transparency and reproducibility, enabling independent verification and replication of our results.

.1 Environment and Tooling

All experiments were performed in a CUDA-enabled JupyterHub environment using Python 3.10 and PyTorch 2.5.1 with CUDA 12.1 support. The SynLLM pipeline was built using the Hugging Face `transformers` library (v4.33.0) for model loading and inference, along with `accelerate` (v1.4.0) for efficient device management and parallel execution if needed. Quantized inference at 4-bit and 8-bit precision was enabled using the `bitsandbytes` library (v0.45.3).

Experiments were executed on a **single NVIDIA L40 GPU** with 48 GB of available GDDR6 VRAM and CUDA driver version 550.127.05, under CUDA runtime 12.4, in a JupyterHub environment.

.2 Model Configuration and Inference

Each large language model (LLM) used in SynLLM was configured and executed in a zero-shot inference setting, with prompt-based control tailored for structured medical data generation. To ensure compatibility with the model's pretraining and tokenization schemes, we dynamically mapped model names to the appropriate chat style (e.g., CHATML, LLAMA, OPENCHAT), and applied model-specific prompt templates at runtime.

Models were loaded using the Hugging Face `tTransformers` library, with quantized 4-bit inference enabled via the `bBitsaAndbBytes` package library. The configuration leveraged `bnb_4bit_quant_type="nf4"` with `"nf4" with float16` computation for memory-efficient deployment. For LLaMA-based architectures, rotary positional embedding scaling (`rope_scaling`) was applied where available to support longer sequence contexts. Models incompatible with quantization were automatically reverted to standard full-precision loading.

At generation time, system and user prompts were formatted using model-specific conventions and tokenized using the model's native tokenizer. Tokenization padding and truncation were configured based on model context window limits, with truncation applied to avoid overflow.

Generation was conducted in mini-batches of 20 using top-p sampling ($p = 0.9$) with temperature 0.7. These parameters were fixed across all models for consistency. Preliminary trials with temperature values in $\{0.5, 0.7, 0.9\}$ revealed minimal qualitative differences in output structure, but higher temperatures increased the likelihood of generating rare or edge-case records. We leave a systematic exploration of sampling hyperparameters to future work. Outputs were parsed line-by-line into structured patient records, and only samples conforming to the expected schema were retained. Invalid generations were logged to a rejection report. The final dataset was written to disk in CSV and JSON format.

System metrics, including GPU memory before and after generation, CPU and RAM usage, and total runtime, were logged per model and prompt.

This inference pipeline allows SynLLM to evaluate a wide range of open-source LLMs in a unified and controlled setting, with minimal memory overhead and consistent record formatting across all prompt-model configurations.

.3 Prompt Templates

This section presents abridged versions of the structured prompt templates employed in SynLLM. While templates are designed to be dataset-agnostic, the examples below reflect their instantiation for the *Diabetes* dataset. At runtime, each prompt is dynamically populated with schema-level information, statistical summaries, and clinical constraints specific to the target dataset. All templates begin with a shared system message that standardizes the generation format:

```
System: Generate k patient records in newline-delimited JSON
format. Do not include any explanation or commentary. Adhere
strictly to the schema and guidelines provided.
```

Prompt A – SeedEx (Minimal Example-Based Prompt)

```
Generate realistic synthetic patient records for diabetes prediction using the
    following structure.

gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level,
    blood_glucose_level, diabetes

Example Records:
Female,45.2,1,0,never,28.5,6.2,140,0
Male,62.7,1,1,former,32.1,7.1,185,1
...
```

Prompt B – FeatDesc (Feature Description Prompt)

```
Generate realistic synthetic patient records for diabetes prediction.

Features:
1. gender: Patient’s gender (Male/Female)
2. age: Age in years (Float: 18.0-80.0)
3. hypertension: Hypertension diagnosis (0: No, 1: Yes)
4. heart_disease: Heart disease diagnosis (0: No, 1: Yes)
5. smoking_history: Smoking status (never/former/current/not current)
6. bmi: Body Mass Index (Float: 15.0-60.0)
7. HbA1c_level: Hemoglobin A1c (Float: 4.0-9.0)
8. blood_glucose_level: Glucose level in mg/dL (Int: 70-300)
9. diabetes: Diabetes diagnosis (0: No, 1: Yes)

Example records:
Female,45.2,1,0,never,28.5,6.2,140,0
Male,62.7,1,1,former,32.1,7.1,185,1
...
```

Prompt C – StatGuide (Metadata-Augmented Prompt)

```
Generate realistic synthetic patient records for diabetes prediction.

Feature Metadata:
gender: Male: 48%, Female: 52%
age: Mean: 41.8, Std: 15.2, Range: 18-80
hypertension: No: 85%, Yes: 15%; correlated with age, BMI
heart_disease: No: 92%, Yes: 8%; correlated with age, hypertension
smoking_history: never: 60%, former: 22%, current: 15%, not current: 3%
bmi: Mean: 27.3, Std: 6.4, Range: 15-60
HbA1c_level: Mean: 5.7, Std: 0.9, Range: 4.0-9.0; correlated with diabetes
glucose: Mean: 138.0, Std: 40.5, Range: 70-300; correlated with HbA1c_level
diabetes: No: 88%, Yes: 12%; correlated with HbA1c_level, glucose

Example records:
Female,45.2,1,0,never,28.5,6.2,140,0
Male,62.7,1,1,former,32.1,7.1,185,1
...
```

Prompt D – ClinRule (Clinically Constrained Prompt)

```
Generate realistic synthetic patient records for diabetes prediction.

Feature Metadata:
gender: Male: 48%, Female: 52%
age: Mean: 41.8, Std: 15.2, Range: 18-80
hypertension: No: 85%, Yes: 15%
```

```
heart_disease: No: 92%, Yes: 8%
smoking_history: never: 60%, former: 22%, current: 15%, not current: 3%
bmi: Mean: 27.3, Std: 6.4, Range: 15-60
HbA1c_level: Mean: 5.7, Std: 0.9, Range: 4.0-9.0
glucose: Mean: 138.0, Std: 40.5, Range: 70-300
diabetes: No: 88%, Yes: 12%

Maintain the following correlations:
- Higher age is associated with hypertension and heart disease
- Higher BMI increases diabetes risk
- HbA1c_level correlates with diabetes
- Glucose correlates with HbA1c_level and diabetes
- Hypertension and heart disease more common with age

Each record must follow:
gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level,
    blood_glucose_level, diabetes
```

.4 Dataset Information

To support reproducibility and deeper analysis, we summarize the datasets used in our experiments. Table 5 provides additional details, including the number of records, total features, and the breakdown by feature type (numerical, categorical, binary). Each dataset is used for a classification task, diabetes diagnosis (binary), stroke prediction (binary), and cirrhosis severity (multi-class). Target columns are preserved during synthetic generation, and standard preprocessing steps (missing value removal, range clipping, encoding) were applied consistently across real and synthetic versions. All datasets are publicly available and commonly used in medical ML research.

Table 5: Summary of datasets used in experiments. Num. = numerical, Cat. = categorical, Bin. = binary features.

Dataset	Records	Features	Num.	Cat.	Bin.	Task Type	Target Column
Diabetes [3]	100,000	9	4	2	3	Binary Classification	`diabetes`
Stroke [2]	5,110	12	4	5	3	Binary Classification	`stroke`
Cirrhosis [1]	418	20	12	8	0	Multi-class Classification	`severity`

.5 LLM Selection

To evaluate prompt-driven synthetic data generation across architectures and tuning styles, we benchmarked SynLLM using 20 open-source language models. These models span various families (e.g., LLaMA, Mistral, Yi, GPT), fine-tuning methods (e.g., instruct, chat, DPO), and context lengths. Table 6 summarizes each model's key properties, including parameter size, fine-tuning strategy, and maximum context window. More recent releases such as Qwen3-8B, DeepSeek variants, and LLaMA 3 13B were not included at the time of evaluation due to timing constraints. We plan to extend SynLLM to incorporate these newer models in future work, along with a longitudinal analysis of LLM evolution and its impact on synthetic data fidelity and privacy.

.6 Machine Learning Utility Result Comprehensive Table

Table 7 provides the full results for all 20 LLMs evaluated across both TSTR (Train on Synthetic, Test on Real) and TRTS (Train on Real, Test on Synthetic) setups. Each model's performance is reported in terms of accuracy, macro-averaged F1 score, and AUC-ROC, averaged across all prompt variants. These results extend the subset summarized in Table 4 in the main paper.

Each model–prompt combination was evaluated across 5 independent runs using different random seeds. We report the mean for all primary metrics to capture performance variability.

Table 6: Evaluated LLMs. FT: Ba=Base, In=Instruct, Ch=Chat, DPO=Direct Preference Optimization, MPT=MosaicML Pretrained Transformer.

ID	Model	Params	FT	Ctx
1	GPT-2 (S/M/L) [28]	0.1–0.8B	Ba	1024
2	Gemma-7B-IT [16]	7B	In	8192
3	InternLM2.5-7B-Chat [9]	7B	Ch	32768
4	LLaMA-2-13B-Chat [35]	13B	Ch	4096
5	LLaMA-2-7B-Chat [35]	7B	Ch	4096
6	LLaMA-3-8B [5]	8B	Ba	8000
7	LLaMA-3.1-8B-Instruct [5]	8B	In	128000
8	Mistral-7B-Instruct [6]	7B	In	32768
9	Mosaic-7B-Instruct [24]	7B	MPT	8192
10	Nous-Hermes-2-Mistral-7B [29]	7B	DPO	32768
11	Nous-Hermes-2-Yi-34B [29]	34B	In	4096
12	OpenChat-3.5-GPTQ [39]	7B	Ch	8192
13	OpenChat-3.5 [32]	7B	Ch	8192
14	Qwen-1.5-7B-Chat [33]	7B	Ch	32768
15	Qwen2-7B-Instruct [34]	7B	In	131072
16	StableBeluga-7B [7]	7B	Ch	4096
17	Yi-6B-Chat [4]	6B	Ch	32768
18	Zephyr-7B-Beta [17]	7B	DPO	32768

Table 7: Model Evaluation: Mean ML Utility Metrics (averaged across all prompts)

Model	TSTR			TRTS		
	Acc.	F1	AUC	Acc.	F1	AUC
GPT-2-Large	0.90	0.50	0.83	0.86	0.81	0.90
GPT-2-Medium	0.92	0.57	0.85	0.88	0.76	0.98
GPT-2-Small	0.90	0.53	0.86	0.93	0.87	**0.99**
Gemma 7B	0.90	0.59	0.87	0.90	0.89	0.94
InternLM2.5 7B	0.89	0.53	0.89	0.89	0.84	0.98
LLaMA 2 13B	0.82	0.54	0.66	0.74	0.58	0.81
LLaMA 2 13B Chat	0.79	0.54	0.85	0.95	0.92	0.96
LLaMA 2 7B	0.81	0.60	0.80	0.92	0.85	0.87
LLaMA 3 8B	0.90	0.56	0.78	0.83	0.73	0.88
LLaMA 3.1 8B	0.92	0.67	0.91	0.90	0.84	0.98
Mistral 7B	0.90	0.60	0.88	0.82	0.77	0.92
Mosaic MPT 7B	0.92	0.55	0.89	0.88	0.78	0.88
Nous Hermes Mistral 7B	0.90	0.71	0.91	0.79	0.72	0.95
Nous Hermes Yi 34B	**0.93**	**0.74**	**0.92**	0.87	0.75	0.94
OpenChat 3.5	0.92	0.70	0.89	0.86	0.71	0.85
OpenChat 3.5 GPTQ	0.86	0.61	0.89	0.83	0.71	0.86
OpenChat 3.5-0106	0.91	0.57	0.91	0.85	0.74	0.94
Qwen2 7B	0.91	0.60	0.91	0.88	0.85	0.95
StableBeluga 7B	0.90	0.51	0.72	0.94	0.73	0.88
Yi 6B	0.82	0.46	0.79	**0.98**	**0.96**	0.98
Zephyr 7B	0.88	0.54	0.82	0.88	0.74	0.89

.7 Hyperparameter Search Space

To optimize each discriminative model in our ML utility evaluation, we used a predefined hyperparameter search space. Table 8 lists the hyperparameters explored for each model, along with their corresponding ranges or categorical options. Search strategies include grid search or Bayesian optimization depending on the model type and evaluation protocol. LogUniform denotes sampling from a log-scale distribution.

Table 8: Discriminative models: hyperparameter search space.

Model	Hyperparameter	Search Space
Random Forest	n_estimators	[50, 300]
	max_depth	[3, 20]
	min_samples_split	[2, 20]
	min_samples_leaf	[1, 10]
	max_features	{sqrt, log2, None}
	bootstrap	{True, False}
	class_weight	{balanced, balanced_subsample, None}
Logistic Regression	penalty	{l1, l2, elasticnet}
	solver	{liblinear, saga, lbfgs} (based on penalty)
	C	LogUniform [1e-4, 1e2]
	l1_ratio	[0.1, 0.9] (if elasticnet)
	max_iter	1000
	class_weight	{balanced, None}
MLP	n_layers	[1, 3]
	hidden_units	[32, 256] per layer
	activation	{relu, tanh}
	solver	{adam, sgd}
	alpha	LogUniform [1e-5, 1e-1]
	learning_rate	{constant, adaptive}
	learning_rate_init	LogUniform [1e-3, 1e-1] (if sgd)
	max_iter	500
XGBoost	n_estimators	[50, 300]
	max_depth	[3, 10]
	learning_rate	LogUniform [0.01, 0.3]
	subsample	[0.6, 1.0]
	colsample_bytree	[0.6, 1.0]
	reg_alpha	LogUniform [1e-8, 10]
	reg_lambda	LogUniform [1e-8, 10]
	scale_pos_weight	[0.1, 10.0]
LightGBM	n_estimators	[50, 300]
	max_depth	[3, 10]
	learning_rate	LogUniform [0.01, 0.3]
	subsample	[0.6, 1.0]
	colsample_bytree	[0.6, 1.0]
	reg_alpha	LogUniform [1e-8, 10]
	reg_lambda	LogUniform [1e-8, 10]
	class_weight	{balanced, None}
	verbose	-1
CatBoost	iterations	[50, 300]
	depth	[3, 10]
	learning_rate	LogUniform [0.01, 0.3]
	l2_leaf_reg	LogUniform [1e-8, 10]
	class_weights	{None, [1,2], [1,3]}
	random_state	42
	verbose	False

.8 Detailed Statistical Analysis

To complement the average performance rankings shown in the main paper, Table 9 provides a comprehensive statistical breakdown for each model and evaluation direction (TSTR and TRTS). For every model–metric combination, we report the mean, standard deviation, minimum, and maximum scores over all experimental runs. These statistics allow us to assess not only central performance tendencies but also the stability and worst-case behavior of each model.

LightGBM and Random Forest demonstrate consistently high mean scores and low variability across metrics and directions. In contrast, MLP exhibits wider performance ranges and higher standard deviations, indicating greater sensitivity to task or initialization. ROC-AUC values are generally more stable across models compared to balanced accuracy or F1, especially under TRTS evaluation.

Table 9: Detailed Statistical Analysis by Model and Direction

Model	**Direction**	**Mean**	**Std Dev**	**Min**	**Max**
Catboost	TRTS (Bal. Acc.)	0.8846	0.0224	0.8345	0.9206
	TRTS (Accuracy)	0.9295	0.0073	0.9163	0.9450
	TRTS (F1-Score)	0.9282	0.0083	0.9115	0.9448
	TRTS (ROC-AUC)	0.9765	0.0019	0.9721	0.9791
Catboost	TSTR (Bal. Acc.)	0.8884	0.0256	0.8200	0.9209
	TSTR (Accuracy)	0.9293	0.0096	0.9038	0.9425
	TSTR (F1-Score)	0.9280	0.0109	0.8980	0.9425
	TSTR (ROC-AUC)	0.9801	0.0017	0.9768	0.9828
Lightgbm	TRTS (Bal. Acc.)	0.9155	0.0144	0.8823	0.9390
	TRTS (Accuracy)	0.9362	0.0063	0.9237	0.9475
	TRTS (F1-Score)	0.9366	0.0064	0.9231	0.9481
	TRTS (ROC-AUC)	0.9767	0.0013	0.9744	0.9789
Lightgbm	TSTR (Bal. Acc.)	0.9055	0.0110	0.8877	0.9266
	TSTR (Accuracy)	0.9298	0.0033	0.9213	0.9375
	TSTR (F1-Score)	0.9299	0.0033	0.9213	0.9375
	TSTR (ROC-AUC)	0.9785	0.0017	0.9744	0.9814
Logistic Regression	TRTS (Bal. Acc.)	0.8012	0.0331	0.7283	0.8399
	TRTS (Accuracy)	0.8381	0.0141	0.8075	0.8588
	TRTS (F1-Score)	0.8417	0.0083	0.8186	0.8549
	TRTS (ROC-AUC)	0.9045	0.0083	0.8806	0.9120
Logistic Regression	TSTR (Bal. Acc.)	0.8117	0.0346	0.7386	0.8506
	TSTR (Accuracy)	0.8332	0.0227	0.7987	0.8675
	TSTR (F1-Score)	0.8373	0.0143	0.8123	0.8598
	TSTR (ROC-AUC)	0.9227	0.0040	0.9048	0.9261
Mlp	TRTS (Bal. Acc.)	0.6628	0.1149	0.4992	0.8258
	TRTS (Accuracy)	0.8168	0.0353	0.7688	0.8638
	TRTS (F1-Score)	0.7821	0.0699	0.6744	0.8612
	TRTS (ROC-AUC)	0.8533	0.0663	0.6334	0.9146
Mlp	TSTR (Bal. Acc.)	0.6459	0.1285	0.4992	0.8464
	TSTR (Accuracy)	0.8105	0.0445	0.7600	0.8800
	TSTR (F1-Score)	0.7635	0.0861	0.6574	0.8743
	TSTR (ROC-AUC)	0.8586	0.0834	0.5404	0.9281
Random Forest	TRTS (Bal. Acc.)	0.9121	0.0173	0.8711	0.9334
	TRTS (Accuracy)	0.9321	0.0039	0.9237	0.9400
	TRTS (F1-Score)	0.9326	0.0038	0.9254	0.9403
	TRTS (ROC-AUC)	0.9709	0.0040	0.9621	0.9769
Random Forest	TSTR (Bal. Acc.)	0.8999	0.0119	0.8754	0.9193
	TSTR (Accuracy)	0.9314	0.0048	0.9225	0.9400
	TSTR (F1-Score)	0.9310	0.0051	0.9210	0.9401
	TSTR (ROC-AUC)	0.9749	0.0021	0.9684	0.9785

.9 Model Performance Ranking

Table 10 summarizes the average balanced accuracy of five discriminative models evaluated across multiple downstream classification tasks. Balanced accuracy was chosen as the primary metric to account for class imbalance commonly found in real-world healthcare datasets. Each model was evaluated over 60 experimental runs, and the reported values include the mean balanced accuracy, standard deviation, and the full range of scores observed.

LightGBM and Random Forest consistently achieved the highest performance, with mean balanced accuracies exceeding 0.90 and relatively low variability, indicating stable generalization across tasks. CatBoost also performed well but showed slightly higher variance. Logistic Regression showed moderate performance with narrower range, while MLP yielded the lowest mean accuracy and the widest range, suggesting higher sensitivity to task-specific or tuning variations.

Table 10: Performance Ranking of Models by Balanced Accuracy

Rank	**Model**	**Mean Acc.**	**Std Dev**	**Range**
1	LightGBM	0.9105	0.0137	[0.8823, 0.9390]
2	Random Forest	0.9060	0.0161	[0.8711, 0.9334]
3	CatBoost	0.8865	0.0241	[0.8200, 0.9209]
4	Logistic Reg	0.8064	0.0342	[0.7283, 0.8506]
5	MLP	0.6544	0.1222	[0.4992, 0.8464]

Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction describe all sections of the paper in a precise manner. Our main contribution is

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the discussion section, we have mentioned all the limitations and potential work for extending the research

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There is no new theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our code is reproducible; however, in some cases, it depends on some conditions, including hardware, network bandwidth, etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the whole history of the code has been recorded and is ready to be provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We mentioned our experiments' documentation in the supplementary section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the details regarding all the metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We quantify and report hardware and time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not believe our dataset has a high potential for misuse.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We did not discuss the impact of our research on society in detail. However, it clearly can help institutions to eliminate the use of real patients' data for different purposes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: All the models and datasets are open-source.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The license was issued in the code-base.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new synthetic datasets and code for medical data generation using LLMs. We provide documentation detailing the generation pipeline, prompt formats, evaluation metrics, and usage instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects research or crowdsourcing. All data used for generation and evaluation are synthetic or publicly available and de-identified, and no experiments were conducted on real individuals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were only used for mundane writing and formatting tasks.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.