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


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.



***Arshia Ilaty,*** *Hossein Shirazi, and Hajar Homayouni*



*This research is supported by the SDSU Presidential Graduate Research Fellowship, the University of CA, Irvine, and the Computational Science Research Center (CSRC) at San Diego State University. This research has benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program*


**Algorithm 1:** SynLLM: Structured Medical Data Generation with LLMs

**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 $\mathcal{R}$;

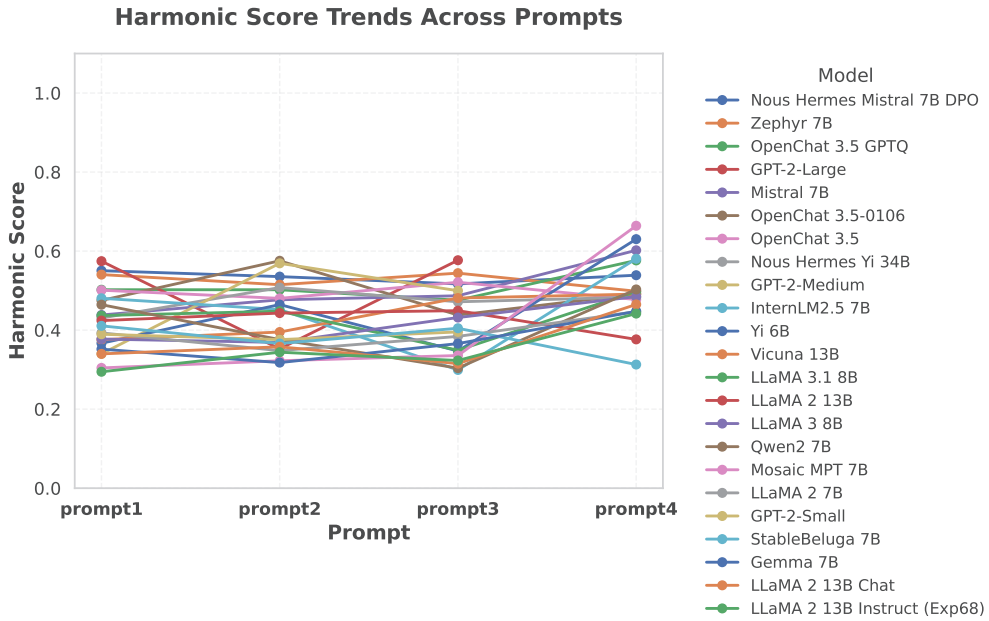
**Stage 3: Synthetic Data Generation**
**foreach** *model* $m \in \mathcal{M}$ **do**
  **foreach** *prompt* $P$ **do**
    Generate synthetic records $R_{m,P} = m(P)$;
    Parse $R_{m,P}$ into structured tabular format;

**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;

**return** $\hat{\mathcal{D}}_{syn} = \bigcup R_{m,P}$

| 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 |

| 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 |




**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 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
```