William Chang Liu - 1009048852

Ryan Zhang - 1009020453

Taiyi Jin - 1009075796

Abdus Shaikh -

CSC490 Assignment 1: Project Landscape

Table of Contents

Part 1: Interest Statements

Our team aims to investigate LLM-based synthetic data generation within the healthcare domain. Data availability is often the first major challenge in developing machine learning models. Data can be expensive, scarce, and/or increasingly regulated—particularly in healthcare. Synthetic data offers a potentially promising way to address these concerns by enabling model training on synthetic datasets that can support real-world applications. However, current approaches remain limited: synthetic datasets frequently diverge from real-world distributions, lack well-established evaluation metrics, pose unresolved questions around privacy guarantees, and struggle to capture deeper correlations in complex data. We chose healthcare as our focus because it embodies these challenges most clearly, from restrictions on EHR access to the difficulty of modeling underlying diagnostic and treatment patterns. Our objective is to explore these challenges, experiment with methods of synthetic data generation, and investigate the potential of synthetic data in replicating real-world patterns and relationships.

Comments from the team members:

William Liu - *I am excited to investigate the pros and cons of different approaches to systematically evaluating synthetic data, and to experiment with new generation methods that could improve its accuracy and reliability.*

Ryan Zhang - *I am looking forward to investigating the efficacy of our data generation models in real-world scenarios, and to see what system design elements we could potentially incorporate to elevate our research from findings to a useful product.*

Abdus Shaikh - *I am excited to see where current state-of-the-art synthetic data generation models fall short and why. I want to know if there is a fundamental pattern to healthcare data that limits the utility of synthetic data, or if there are perspectives that haven't been considered yet that would enable effective synthetic data generation.*

Taiyi Jin - *I am eager to investigate whether our improved synthetic data generation method could yield something that not only protects privacy but also retains the insights needed for training effective models, and hope that our work can contribute to building trustworthy synthetic data pipelines.*

Part 2: Landscape Analysis

Relevant Item	**Description**	**Commentary**
Meditron	A suite of open-source LLMs adapted from Llama-2 by further pre-training on curated medical data.	Could serve as a good base model to tune with real and synthetic MHRs. Models focus on differential diagnosis.
MedBERT	A transformer-based language model pretrained on large-scale biomedical and clinical text, including electronic health records (EHRs)	One of the strongest advantages of MedBERT is that it is fully open source — both the pretrained weights and the training scripts are publicly available. It also has a paper written on it. It can be a good starting point to compare performance of its current implementation to one using our synthetic data to fine tune instead.
BioMistral	A suite of open-source LLMs adapted from Mistral by further pre-training on curated medical data.	Biochem research oriented. "Best performing medical LLM in this (7B) weight category"
K. Dankar, I. Mahmoud. Fake it till	A research paper investigating how different data-generation	Some very useful guidelines here on how to make effective

you make it: guidelines for effective synthetic data generation, Applied Sciences 11.5 (2021): 2158, 2021.	settings affect the utility of synthetic datasets. Also dives into ways to measure the utility of synthetic datasets, specifically propensity and how it correlates with model accuracy.	synthetic data models. Established propensity as a "the most practical measure for predicting the overall utility of a synthetic dataset". Dataset propensity may be a good heuristic for us to follow.
Bt-GAN	A GAN that explicitly tries to reduce bias and preserve fairness in synthetic EHR data. It includes mechanisms for preserving subgroup distributions and reducing illegitimate correlations.	Rather than just looking realistic at the surface, Bt-GAN focuses on preserving subgroup distributions and avoiding distorting underrepresented groups. We can do a case study on how this is achieved
Gretel.ai	A development platform for synthetic data generation. Provides Data Generation, Data Transformation, Data Quality Evaluation, API Integration.	Models use GANS or RNNs. Features seem to be similar to what we want to implement.
T. E. Raghunathan, Synthetic data, Annual Review of Statistics and Its Application, 8, 129-140, 2021.	An overview of the field of synthetic data, going over mathematical and statistical concepts as well as some applications.	The paper touches on the basic idea and math behind differential privacy, which is something we will need to prove our models generate private data.
Making ML models differentially private: Best practices and open challenges	Short blog from Google summarizing established methods of making differentiably-private machine learning models. Methods of combating utility loss. As well as areas of needed improvement in the DP-ML field.	Incredibly practical blog that offers solutions to problems we will likely face in trying to make our model differentiably-private. Suggests methods like DP-SGD and DP-FTRL as industry-standard. Also outlines 4 areas of improvement in the DP-ML field, which we can directly research.
https://github.com/statice/awesome-synthetic-data	Curated list of Synthetic Data Tools.	Gives us a list of libraries we could use to implement our model. Also has an entire list

		of private tools that we can use as inspiration.
Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains: https://par.nsf.gov/servlets/purl/10498692	The authors introduce key concepts and metrics for evaluating LLM-based sampling, including various sampling methodologies and prompting strategies. The paper identifies challenges in achieving desired output distributions.	Gives us insights on how reliable LLMs currently are when it comes to replicating distributions, and the need for careful consideration of sampling methods and prompt design.
CTGAN	A deep learning-based synthetic data generator designed specifically for tabular data. It handles mixed data types (continuous and categorical) and imbalanced distributions by using a conditional generator that models the probability of each column given the others	CTGAN is an interesting case study because it learns the joint distribution of all features in the dataset, allowing it to generate realistic synthetic rows that preserve correlations and dependencies.

Part 3: Project Outline

Problem statement:

The effectiveness of machine learning models in healthcare is often hindered by the limited availability of high-quality training data. Access to real-world healthcare records is restricted by its collection cost and rigorous data privacy laws. Researchers have proposed synthetic data generation methods as a solution, creating the possibility of training models without directly relying on hard-to-access and sensitive patient information. However, current methods are limited in their ability to mimic real-world distributions while respecting differentiable privacy concerns. These gaps limit our ability to create useful machine learning models for healthcare and address problems that can only be addressed by such models. We aim to close these gaps and create robust synthetic data generation methods that respect privacy while also enabling large scale and effective model training.

A Proposed solution

We will create a fine-tuned LLM that can generate hundreds to thousands of synthetic healthcare records that do not reveal any private information about the real-world patients in its training data. This LLM will be highly specialized to maximize its effectiveness, meaning it will not be conversational in nature as typical LLMs, rather, it will consume certain requirements such as geographical location, age groups, demographics, and generate synthetic health records that accurately reflect these parameters while respecting the privacy of the real-world patients in its training and fine-tuning data.

A High-level technical approach and milestones

We will begin by experimenting with and evaluating state-of-the-art synthetic data generation models. We will focus on LLMs fine-tuned on healthcare data, but we will also take inspiration from synthetic data generators for other domains where we see fit.

We break down our high-level milestones like so

1. Assess and evaluate current state-of-the-art methods. Where do they fall short? What do they do well? What novel techniques did they use?
2. Establish industry standard benchmark metrics for our domain for use as our guide during development. We will aim to Improve all of the benchmarks we establish and consider doing so as project success.
3. Develop on our novel approach. This will likely consist of taking the state-of-the-art model and experimenting with it. We will fine-tune it, add and/or remove modules, combine it with other SOTA methods, and evaluate our findings as we go to inform our decisions

4. Compare against established benchmarks. Once we have established that our model outperforms other SOTA models in most if not all metrics, we will consider our approach as successful
5. Write a report with our findings, techniques, model limitations, and next steps.

A list of unknowns to investigate

1. Where do current synthetic data generation models fall short? Is this a fundamental problem with the domain or can novel approaches to model design address the current shortcomings?
2. How can we maximize data privacy while also maximizing data utility?
3. What makes synthetic data different from real-world data? Can we close this gap?

Part 4: Project Press Release

HealthForge: Unleashing the Potential of Synthetic Healthcare Data

Healthcare AI research has long been limited by the scarcity of accessible, high-quality patient data. Real-world electronic health records (EHRs) are costly, sensitive, and tightly regulated, making it difficult for researchers and developers to train machine learning models that can truly improve outcomes.

Today, our research team is proud to announce **HealthForge**, a synthetic healthcare data generator powered by fine-tuned large language models (LLMs). HealthForge creates **realistic yet privacy-preserving synthetic healthcare records**, enabling researchers and developers to train, test, and evaluate AI systems without exposing sensitive patient information.

Unlike other synthetic data platforms, HealthForge is **built for healthcare**. It can generate patient records tailored to parameters such as demographics, geography, and age groups, while maintaining the complex correlations between diagnoses, treatments, and outcomes. Our approach incorporates **privacy safeguards** like differential privacy, ensuring that generated data cannot be traced back to real patients.

Why HealthForge

- **Privacy preserving**: Protects patient confidentiality while retaining predictive patterns critical for training models.
- **Customizable datasets**: Generate healthcare records for specific research scenarios, such as diabetes, ICU stays, or geriatric care.

- **Research level quality**: Designed to benchmark against real-world data, providing utility for downstream machine learning tasks.
- **Scalable innovation**: Enables institutions of any size to test healthcare AI models without requiring direct access to sensitive hospital records.

Voices from the Field

- *"HealthForge has transformed the way we experiment with predictive models. I no longer have to worry about patient privacy when testing new algorithms."* — **Dr. Maria Lopez, Clinical Researcher**
- *"We needed a safe way to share EHR data with collaborators. HealthForge allows us to do that without risking compliance or patient trust."* — **Oliver Wang, Health IT Manager**
- *"Using HealthForge for my ML project was eye-opening—the synthetic data preserved real-world patterns so well, I could focus on model development rather than data collection hurdles."* — **Priya Nair, Graduate Student**

HealthForge represents a step toward making healthcare AI research more accessible, ethical, and impactful. By bridging the gap between **data scarcity** and **data privacy**, we are creating a foundation for safer, smarter, and more trustworthy healthcare technologies.

Part 5: Appendix

Project Press Release Iterations:

Iteration 1

Privacy focused, focusing on protecting patient confidentiality as the main selling point. Highlight the privacy guarantees and trust for hospitals.

Key Angle: Synthetic data is valuable only if privacy is protected. Our contribution is ensuring patient confidentiality without compromising research.

Target Audience: hospital administrators, policymakers

Iteration 2

Tech innovation focused, focusing on the use of LLMs and the advanced synthetic data generation methods. Show the novelty of LLMs in the field of EHR.

Key Angle: Our synthetic generator is not just another data tool — it's powered by cutting-edge AI (LLMs fine-tuned on healthcare data) that sets a new benchmark in synthetic data generation.

Target Audience: ML researchers, data scientists

Iteration 3

Future potential focused, focusing on the scalability of the usage of synthetic healthcare data with our method.

Key Angle: Synthetic data isn't just a workaround — it's the future infrastructure of collaborative healthcare AI research.

Target Audience: investors, healthcare innovation leaders

Iteration 4 (current version)

Real-world impact focused, while taking a more balanced approach in terms of the project introduction, focusing on how people use synthetic data to train models and advance healthcare AI.

Key Angle: Our project will be useful today and be relevant to everyone in the field.

Target Audience: general academic and professional readers