SynGen: Synthetic Data Generation Leveraging Large Language Models for Patient-Trial Matching.

Enhancing Privacy in Clinical Text Mining with

Specific Aims

The healthcare industry generates vast amounts of data, including sensitive patient information, which has led to the emergence of clinical text mining as a promising tool for extracting valuable insights from unstructured data. This, in turn, improves patient outcomes and reduces healthcare costs (Denny et al., 2013). According to recent market research, the global clinical data analytics market is projected to reach $22.5 billion by 2025, with a compound annual growth rate (CAGR) of 12.2% (MarketsandMarkets, 2020). This highlights the increasing importance of finding solutions that balance data privacy with extracting valuable insights from clinical text mining. Large language models (LLMs) have emerged as a particularly effective approach to clinical text mining (Zhang et al., 2020). However, their use raises significant privacy concerns, as they can extract sensitive information from electronic health records (EHRs). Ensuring patient privacy and compliance with healthcare regulations is crucial, especially since even anonymized data is vulnerable to re-identification through de-anonymization (Chen et al., 2021; Shickel et al., 2017). Current commercial solutions in the clinical text mining domain focus on extracting valuable insights from unstructured text data. These solutions often leverage advanced natural language processing and machine learning algorithms and offer various tools for data exploration, analysis, and synthetic data generation to accommodate different research needs ^[1]. However, there are some limitations to the current commercial solutions. **(1)** Data privacy remains a major concern, as these platforms must constantly balance the extraction of valuable insights with ensuring the protection of sensitive patient information. This challenge is exacerbated by the risk of re-identification through de-anonymization ^[2]. **(2)** Additionally, while many of these solutions offer pre-built models for various clinical use cases, they may not be suitable for researchers who require a high level of customization. In such cases, the pre-built models may not be easily adaptable and might not address researchers' specific research needs. **(3)** furthermore, some platforms are difficult to navigate for users without technical expertise due to their advanced features, creating a barrier to entry. ^[3].

In response to the limitations of current commercial solutions, we propose SynGen, an online-based platform designed to address the following three major concerns in clinical text mining with synthetic data generation using LLMs: **(1) Data Privacy**: SynGen will enhance privacy by reducing the need for uploading patient data to public tools such as ChatGPT, effectively mitigating privacy concerns. By utilizing synthetic healthcare data generation, the platform will create data that simulates real-world clinical data without exposing sensitive patient information. **(2) Level of Customization**: SynGen is designed to be highly customizable, allowing users to tailor the synthetic data generation process to their specific research needs. This flexibility will ensure the platform can accommodate various research requirements, addressing the limitations of pre-built models in other commercial solutions. **(3) Ease of Use**: To ensure user-friendliness, SynGen will employ text-based natural language processing technologies and offer a comprehensive training paradigm that enables users to learn how to generate, explore, and analyze synthetic data effectively. This approach will make the platform accessible to users with varying levels of technical expertise, facilitating seamless adoption and integration of SynGen's capabilities. With these features, SynGen aims to transform the healthcare industry by enabling more efficient and secure data analysis, contributing to improved patient outcomes through better insights derived from synthetic data, and minimizing costs associated with data collection and processing by providing a robust, customizable solution.

In Phase, I, the AI POW LLC team, and UTHealth propose to develop SynGen, a platform focusing on biological named entity recognition and relation extraction. The goal is to assist researchers and data scientists create large datasets for machine learning and AI-based applications. During this phase, the team will also determine the feasibility and user acceptability of SynGen, which is designed to enhance privacy in clinical text mining through synthetic data generation using large language models.

Aim 1: Data synthesis and augmentation with generic LLM - NER, RE.

Aim 2: Patient-clinical trial matching.

Aim 3: Analytical validation of patient-trial matching prototype.

SW development. Locally trained LLM for [target application] - human-in-the-loop data verification

In **Aim 1: Enhancing Data Privacy with Advanced Algorithms**, our team will refine algorithms and natural language processing techniques to generate secure and privacy-preserving synthetic data. In **Aim 2: Offering High-Level Customization,** we will address the need for a high level of customization by developing a highly customizable platform, allowing users to tailor the synthetic data generation process to their specific research needs. Finally, in **Aim 3: Ensuring Ease of Use and Accessibility for Users**, we will employ text-based natural language

processing technologies and offer a comprehensive training paradigm that enables users with varying levels of technical expertise to effectively generate, explore, and analyze synthetic data, thereby facilitating seamless adoption and integration of SynGen's capabilities.

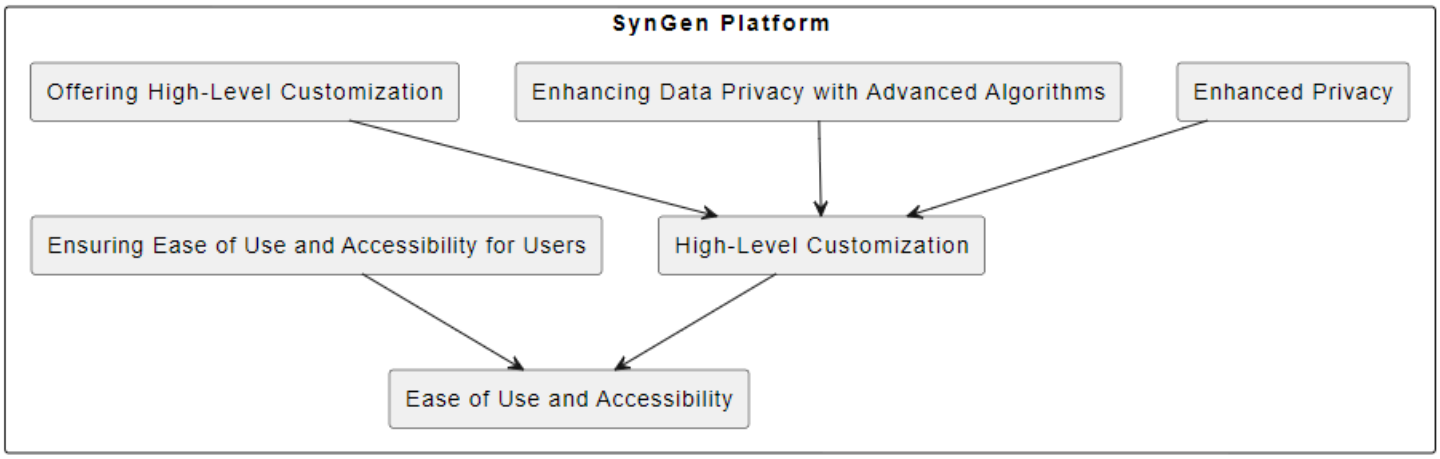
At the onset of Phase 2, SynGen will revolutionize the field of clinical text mining by refining algorithms and natural language processing techniques for generating synthetic data and enhancing the accuracy of the generated data. By doing so, **SynGen will enable clinicians and researchers to analyze and understand better and healthcare data, ultimately leading to improved patient outcomes as healthcare providers become better equipped with easily digestible information to make informed decisions**. Furthermore, SynGen's ease-of-use and customization options will streamline the data generation and analysis process, allowing providers to use their time and resources more efficiently and **reducing healthcare costs**.

Research Strategy
(A) Significance

Clinical text mining has become increasingly important for healthcare researchers and clinicians to extract valuable insights from vast amounts of unstructured clinical data. According to a recent report by Grand View Research, the global clinical data analytics market is expected to reach over USD 16 billion by 2025, with clinical text mining contributing to this growth.^[1.A] However, privacy concerns often limit the use of such data due to ethical and legal challenges associated with real patient data. The Health Insurance Portability and Accountability Act (HIPAA) in the United States and the European Union's General Data Protection Regulation (GDPR) are just two examples of the regulations that govern the use of patient data. Consequently, there is a growing demand for solutions that enable researchers to access clinical data while preserving patient privacy.

One solution to this challenge is using synthetic data generated from large language models. Synthetic data can be used to train machine learning models without the risk of exposing sensitive patient information, making it a promising approach to preserving patient privacy in clinical text mining. Current approaches to clinical text mining, such as natural language processing and machine learning algorithms, can extract valuable insights from unstructured clinical data, which can improve patient outcomes and inform healthcare policy. Additionally, synthetic data can help to overcome the limitations of small sample sizes and data imbalance that are common in healthcare research.^[2.A]. However, privacy concerns remain a major challenge in using such data, as sensitive patient information is vulnerable to re-identification through de-anonymization (Chen et al., 2021; Shickel et al., 2017). Additionally, some platforms may be challenging to navigate for users without technical expertise, creating a barrier to entry. Finally, while many solutions offer pre-built models for various clinical use cases, these models may not be customizable enough to meet researchers' specific needs.^[1]

To address this challenge, we introduce a novel approach to enhancing privacy in clinical text mining that utilizes synthetic data generation with large language models. SynGen, our proposed innovative system, employs state-of-the-art algorithms and natural language processing techniques to generate synthetic data that accurately reflects real-world patient data. It also features a user-friendly interface and customizable options, allowing healthcare providers to tailor the generated data to their specific needs. The development of this approach is a collaborative effort between healthcare researchers and AI experts, led by AI POW LLC, a Texas-based company that integrates automation and interpretability technologies for Machine Learning (ML) to help researchers better understand the data they generate. AI POW LLC's expertise in automation and interpretability technologies for ML ensures that the proposed system is not only accurate and efficient but also transparent and interpretable. The integration of AI POW LLC's innovative approach to clinical text mining has the potential to revolutionize the way that healthcare data is generated and analyzed (1.1). By providing a more accurate and efficient means of analyzing clinical data while also protecting patient privacy, the proposed system has the potential to improve healthcare outcomes and reduce healthcare costs. As such, it represents a major step forward in the ongoing effort to harness the power of AI and machine learning to benefit patients and healthcare providers alike. The following sections will outline the innovation behind the proposed system. The approach used to develop SynGen and its potential impact on healthcare outcomes and costs.



(B) Innovation:

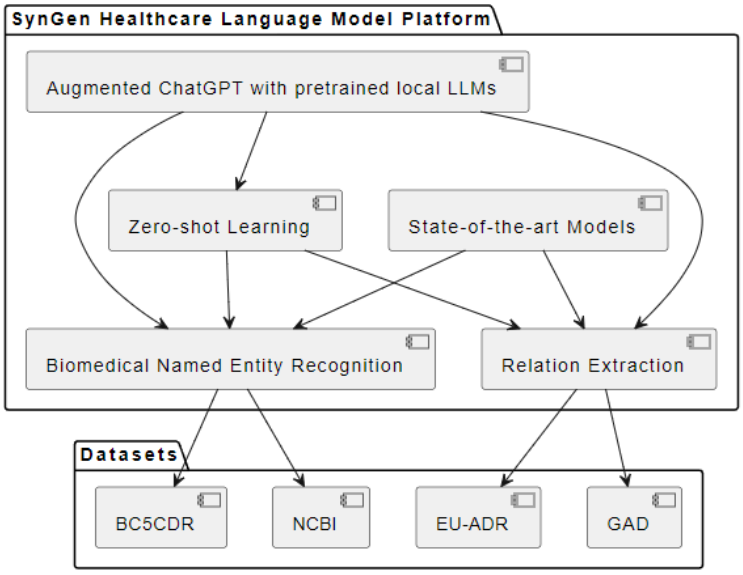
The current model offers a comprehensive and advanced solution for clinical text mining that outperforms existing approaches through several innovative aspects. First, it provides robust privacy protection by generating synthetic data that closely mimic real patient data, effectively reducing the risk of re-identification compared to traditional de-identification methods. Second, the model employs state-of-the-art algorithms and large language models to create high-quality

synthetic data, ensuring an accurate representation of original datasets and overcoming the limitations of traditional methods that may fail to capture the complexity and nuances of clinical data. Third, the user-friendly interface caters to users with varying levels of technical expertise, promoting seamless adoption and integration of the system while breaking down barriers created by solutions with steep learning curves. Finally, the model's flexible platform and numerous customization options empower users to tailor the synthetic data generation process to their specific research needs, enhancing its applicability across a diverse range of research scenarios. These innovative features collectively contribute to a groundbreaking solution that revolutionizes the field of clinical text mining.

SynGen aims to offer a comprehensive and advanced solution for clinical text mining that leverages innovative technologies to specifically improve privacy, ease of use, and customization. By incorporating differential privacy [1.i], a privacy-preserving technique that adds controlled noise to the data generation process, the system ensures robust privacy protection for sensitive clinical data. Advanced natural language understanding (NLU) and generation (NLG) techniques [2.i.] are integrated with large language models to enable better processing of clinical text data and generate high-quality synthetic data that closely resembles the original dataset, enhancing the ease of use through automation and accurate data generation. Furthermore, our team proposes configurable synthetic data generation by providing a range of configurable settings for synthetic data generation, such as the desired level of privacy protection, data distribution, and statistical properties. This will enable users to generate synthetic data that closely align with their research objectives and constraints.

(C) Approach
Preliminary studies

In response to the urgent need for a new approach to language models in healthcare, we conducted a series of experiments using ChatGPT to extract structured information from unstructured healthcare texts. We focused on two critical tasks: Biomedical Named Entity Recognition (NER) and Relation Extraction (RE) in medical texts. NER identifies and categorizes medical entities, while RE extracts relationships between them. The NER task employed an IOB tagging scheme, and the RE task was formulated as a classification problem. We used widely recognized datasets, NCBI and BC5CDR for NER, and GAD and EU-ADR for RE, to evaluate the performance of these tasks.



Due to the weakly supervised nature of the GAD and EU-ADR datasets, we enlisted three annotators to manually label 200 data samples from the original test datasets to ensure accurate evaluation. However, our preliminary findings revealed that ChatGPT alone is ineffective for healthcare tasks, underperforming compared to state-of-the-art models trained on the same datasets. This emphasized the need for a language model specifically trained for the complexities of healthcare and the associated privacy implications.

Additionally, our study introduces the concept of zero-shot learning, which empowers LLMs to perform tasks they have not been explicitly trained on. Specifically, we focused on biological named entity recognition (NER) and relation extraction (RE) tasks. However, our preliminary findings revealed a startling truth: despite its impressive capabilities in classic natural language understanding tasks, ChatGPT alone is ineffective for healthcare tasks. Directly using ChatGPT resulted in poor performance when compared to state-of-the-art models trained on the same dataset, as indicated in Tables 1 and 2. This highlights the urgent need for a language model specifically trained for the complexities of healthcare. It is clear that the limitations of current language models in healthcare are not just performance-related, but also have significant privacy implications. Integrating LLMs into hospital systems raises concerns about the confidentiality of patient information, since most LLMs are only accessible through their APIs.

Table 1: NER Test Results of ChatGPT versus SOTA Metrics

Dataset	Metrics	SOTA	ChatGPT	Decrease

NCBI Disease	P	82.87	32.84 ↓	60.27%
	R	89.54	44.86 ↓	49.91%
	F	86.08	37.92 ↓	55.94%
BC5CDR Chemical	P	91.07	5.76 ↓	94.36%
	R	92.24	11.69 ↓	87.31%
	F	91.65	7.72 ↓	91.58%

Table 3: RE Test Results of ChatGPT versus SOTA Metrics

Dataset	Metrics	SOTA	ChatGPT	Decrease
GAD	P	84.28	76.32 ↓	7.96%
	R	94.21	79.82 ↓	14.39%
	F	88.96	78.03 ↓	10.93%
EU-ADR	P	75.81	72.01 ↓	3.80%
	R	81.20	75.43 ↓	5.77%
	F	78.41	73.68 ↓	4.73%

The development of our innovative training paradigm has been a significant accomplishment in utilizing ChatGPT for healthcare tasks. Through our preliminary work, we have discovered that generating synthetic data with labels using ChatGPT and fine-tuning a local pre-trained language model significantly enhances the performance of the local model compared to just using ChatGPT. This approach has allowed us to address the challenges of utilizing ChatGPT for healthcare tasks while also mitigating potential privacy concerns and reducing the reliance on expensive and time-consuming data collection and labeling.

Our experiments on four representative datasets have provided empirical evidence that our pipeline significantly improves the performance of the local model. We have shown that our approach is effective in generating a variety of examples with varying sentence structures and linguistic patterns, which enhances the quality and diversity of the synthetic data. By eliminating low-quality or duplicated samples created by ChatGPT using a post-processing step, we ensure the accuracy and reliability of the synthetic data. Moreover, our innovative training framework has effectively addressed data privacy concerns by reducing the need for uploading patient data to ChatGPT APIs. This is a crucial advantage, as data privacy is a significant concern in healthcare tasks. Our approach reduces the amount of sensitive patient data that needs to be shared, thereby increasing patient privacy and security.

Our preliminary work has demonstrated that our innovative training paradigm is a promising approach to utilizing advanced language models for healthcare tasks. Our research sets the stage for further investigation and development in this field, and we are excited about the potential for our approach to improving healthcare outcomes and patient care.

Phase I Work Plan:

Aim 1. **Enhancing Data Privacy with Advanced Algorithms**.

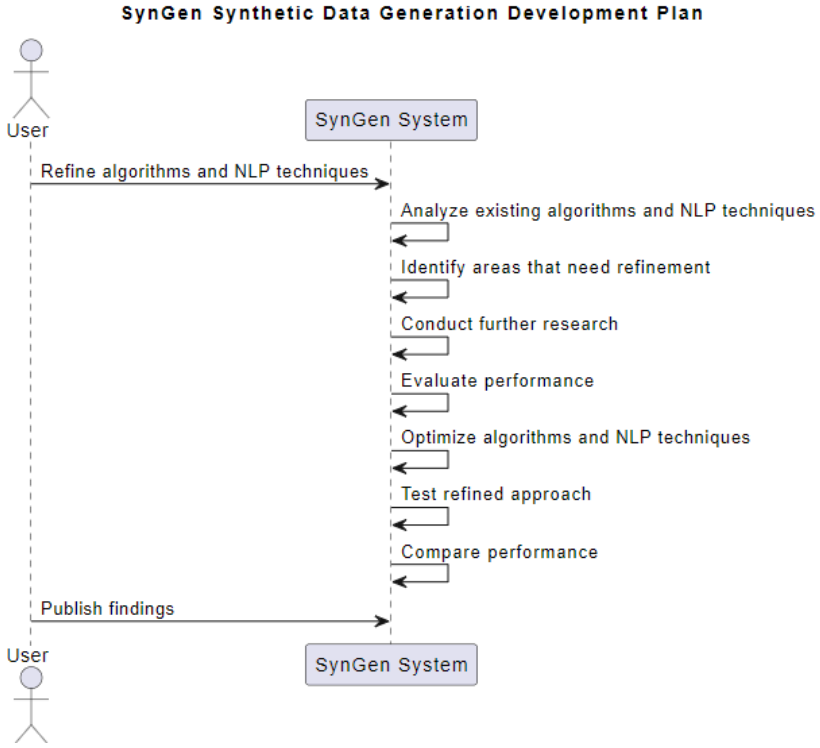
Generating synthetic data is a critical process in the development of advanced language models for healthcare tasks. The first aim of our development plan is to refine the algorithms and natural language processing techniques used to generate synthetic data. In this plan, we aim to evaluate the existing algorithm and NLP techniques, identify areas that need

refinement, conduct further research, evaluate performance, optimize the algorithm and NLP techniques, test the refined approach, compare performance, and publish the findings.

To refine the algorithms and NLP techniques, the first step is to analyze the existing algorithms and NLP techniques used in the preliminary work. Through this analysis, we can identify the strengths and weaknesses of the current approach. Based on the analysis, we can identify the areas that need refinement to improve the performance of the algorithms and NLP techniques. This process will help us refine the algorithms and NLP techniques to enhance the accuracy and reliability of the synthetic data generated.

The next step in our development plan is to conduct further research to refine the algorithms and NLP techniques. This may involve exploring new algorithms or NLP techniques that can improve performance. Once the research is completed, we will evaluate the performance of the refined algorithms and NLP techniques using a range of metrics, including accuracy, precision, and recall. Based on the evaluation results, we will optimize the algorithms and NLP techniques to improve their performance.

Finally, we will test the refined algorithms and NLP techniques to ensure their effectiveness in various contexts. We will compare their performance with the original approach to determine the extent of improvement. The findings of this research will be published in academic journals and presented at conferences to share the knowledge gained through the process. By following this development plan, we aim to refine the algorithms and NLP techniques used to generate synthetic data, which will improve the accuracy and reliability of the data and provide new opportunities for improving healthcare outcomes and patient care.



<u>Aim 1</u> Milestone, Potential Pitfalls, and Alternative Strategies: Milestone: The development plan for refining algorithms and natural language processing techniques in generating synthetic data entails evaluating existing approaches, identifying areas of improvement, conducting research, optimizing techniques, testing the refined approach, comparing performance, and publishing findings. Successfully achieving these milestones will lead to enhanced accuracy and reliability of synthetic data for healthcare tasks.

Potential Pitfalls: Some potential pitfalls to this development plan include the possibility of overfitting, limited availability of diverse real-world data for comparison, inadequate evaluation metrics, difficulties in identifying and addressing biases in the generated data, and potential resistance to adopting synthetic data in healthcare settings due to ethical or privacy concerns.

Alternative Strategies: In order to mitigate potential pitfalls, several alternative strategies can be employed. These may include adopting transfer learning and other techniques to avoid overfitting, partnering with healthcare institutions to access diverse real-world data, utilizing a combination of evaluation metrics to ensure a comprehensive assessment, implementing methods to identify and mitigate biases, and engaging in open communication and collaboration with stakeholders to address ethical and privacy concerns. By considering these alternative strategies, the development plan can be more robust and better prepared to face challenges in refining algorithms and NLP techniques for synthetic data generation.

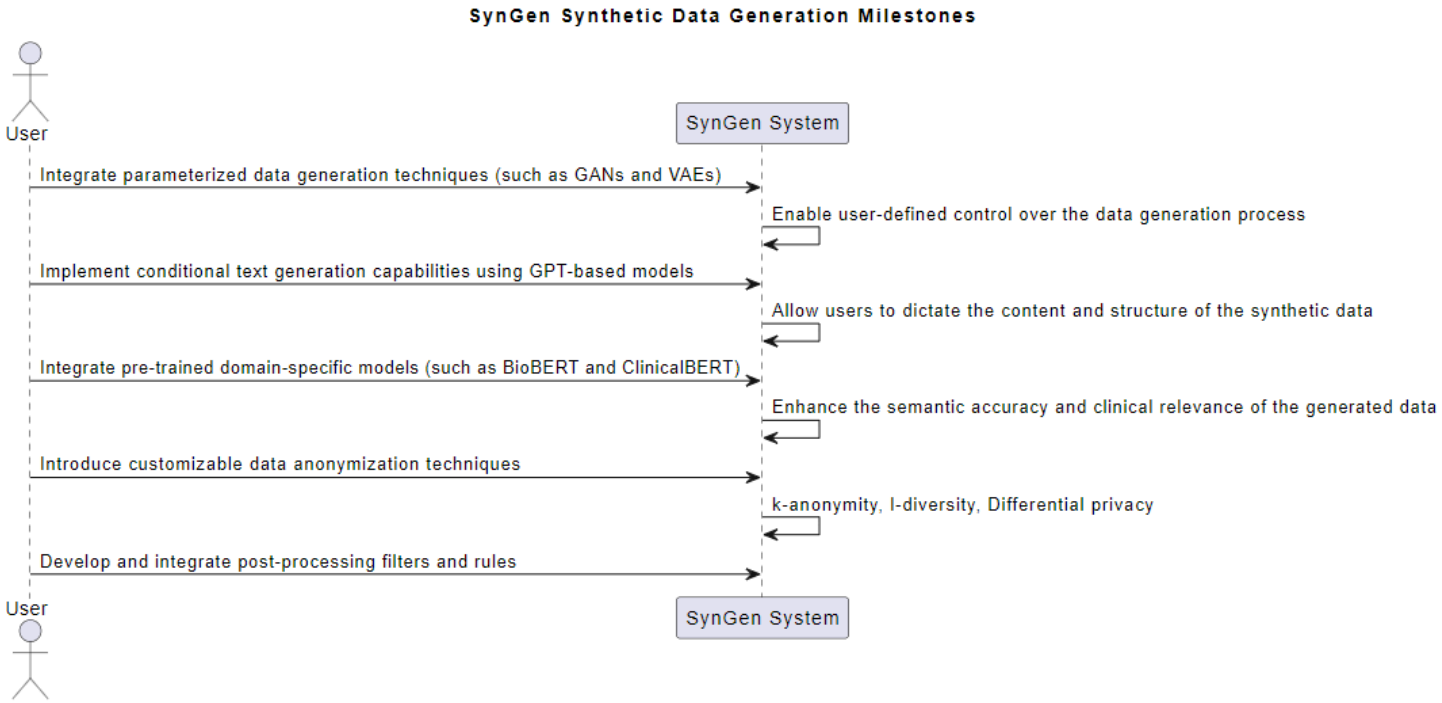
Aim 2. **Offering High-Level Customization:**

The configurable synthetic data generation feature in SynGen aims to give users greater control over the generation process, enabling them to create synthetic data tailored to their specific needs and requirements. To achieve this, we propose the integration of several technologies and techniques. Firstly, parameterized data generation techniques allow users to specify parameters such as the desired distribution of demographic data or the prevalence of specific medical conditions. This can be achieved using techniques like Generative Adversarial Networks (GANs) or Variational

Autoencoders (VAEs), which are capable of generating data with user-defined characteristics. Secondly, we propose the utilization of large language models capable of conditional text generation, such as GPT-based models, to generate synthetic data based on user-defined prompts, constraints, or context. This enables users to control the content and structure of the generated data, making it more relevant and useful for their specific research objectives. Thirdly, pre-trained domain-specific models such as LLaMa can be integrated to ensure that the generated synthetic data is semantically accurate and clinically relevant. These models can capture the specialized vocabulary and linguistic patterns used in the healthcare domain, leading to more realistic and contextually appropriate synthetic data. Fourthly, a range of customizable anonymization techniques such as k-anonymity, l-diversity, and differential privacy can be offered, allowing users to balance the trade-offs between data utility and privacy, depending on their specific requirements and risk tolerance. Finally, a library of customizable post-processing filters and rules can be developed that users can apply to the generated synthetic data to further refine its content, quality, and representativeness. Examples of such filters include removing or modifying specific entities, enforcing constraints on generated relationships, or ensuring that certain patterns or characteristics are preserved in the synthetic data. By integrating these technologies and techniques, SynGen can empower users to generate highly customized synthetic data that meet their specific needs while maintaining data privacy and adhering to regulatory requirements.

<u>Aim 2</u> Milestones, Potential Pitfalls, and Alternative Strategies:

Milestones: Initially, the focus will be on integrating parameterized data generation techniques, such as GANs and VAEs, to enable user-defined control over the data generation process. Next, conditional text generation capabilities using GPT-based models will be implemented to allow users to dictate the content and structure of the synthetic data. In parallel, the integration of pre-trained domain-specific models, such as GPT-based LLaMa, will enhance the semantic accuracy and clinical relevance of the generated data. Subsequently, customizable data anonymization techniques will be introduced, followed by the development and integration of post-processing filters and rules.



Potential pitfalls in this work plan may include challenges in integrating multiple technologies, difficulties in achieving a balance between data privacy and utility, and ensuring that the generated synthetic data is highly quality and representative of the original data. To address these challenges, it is crucial to conduct rigorous testing and validation throughout the development process, collaborate with domain experts, and continuously seek user feedback to refine and improve the system.

Alternative strategies that could be employed in case of difficulties in implementing the proposed work plan include exploring other data generation methods, such as rule-based approaches or template-based generation, or using a combination of pre-processing techniques to improve the input data for language models. Additionally, if integrating specific anonymization techniques proves too complex, alternative privacy-preserving methods, such as secure multiparty computation or homomorphic encryption, can be explored. By being prepared to adapt and modify the work plan based on

encountered challenges, the development of the SynGen system can remain agile and responsive to the needs of the clinical text mining community.

Aim 3: Ensuring Ease of use for Users.

To ensure ease of use for users of SynGen, several technologies can be employed to create a user-friendly and intuitive experience. First, the development of a graphical user interface (GUI) that allows users to interact with the system using visual elements such as buttons, sliders, and dropdown menus can simplify the configuration and execution of data generation tasks. Popular GUI frameworks such as Qt, GTK+, or Electron can be utilized to build the interface. Next, incorporating advanced natural language understanding (NLU) and generation (NLG) techniques, such as those provided by OpenAI's GPT-based models, can enable users to communicate their requirements to the system using natural language commands, making it more accessible for non-expert users. Additionally, embedding tooltips, help menus, and context-sensitive guidance into the interface can provide users with quick and easily accessible information about the system's features and functionalities, further enhancing the overall user experience. Lastly, integrating SynGen with widely-used data analysis platforms and tools, such as Jupyter Notebooks, RStudio, or KNIME, can facilitate seamless adoption and use of the system by researchers and analysts who are already familiar with these platforms.

By combining these technologies, SynGen can provide an intuitive, user-friendly experience that accommodates users with varying levels of expertise, ultimately making the process of clinical text mining with synthetic data generation more accessible and efficient.

Aim 3 Milestone, Potential Pitfalls, and Alternative Strategies:

Milestones: The initial milestone could be the completion of a comprehensive requirement analysis, followed by the selection and integration of appropriate GUI frameworks and NLU/NLG technologies. Subsequently, the development and testing of the interface, along with the integration of tooltips, help menus, and context-sensitive guidance, should be completed. The final milestone would be the integration of SynGen with widely-used data analysis platforms and tools and a successful user testing phase.

Potential Pitfalls: One potential pitfall could be the limited compatibility between chosen GUI frameworks and NLU/NLG technologies, which may cause integration difficulties or impact the overall user experience. Another pitfall could be the difficulty for users to understand and utilize advanced features within SynGen.

Alternative Strategies: To mitigate the risk of limited compatibility, an in-depth evaluation of compatibility during the requirement analysis phase should be carried out. If needed, alternative technologies or frameworks can be considered to ensure seamless integration. To address user experience challenges, extensive usability testing with target user groups, followed by iterations based on user feedback, should be implemented. If user experience remains a challenge, alternative strategies such as developing training materials, video tutorials, or providing live chat support can be employed to enhance user understanding and satisfaction.

Summary: At the end of Phase 1, our team will have implemented several key components of our development plan. These include the creation of an algorithm for data synthesis, testing it, and refining it based on the results of the tests. Additionally, we will have conducted user research, identified user needs, and developed a plan for usability testing. This will set the foundation for the development of a user-friendly interface with customization options for our proposed SynGen. Overall, the successful implementation of these components in our preliminary work will ensure that we are on track to achieve our goals and deliver a comprehensive and effective system. At the end of Phase 2, SynGen will have shown that it can outperform human teams in data analysis and decision-making for complex clinical problems.

Commercialization: Enhancing Privacy in Clinical Text Mining with Synthetic Data Generation using Large Language Models is a promising application in the healthcare industry. The global healthcare artificial intelligence market was valued at $2.1 billion in 2018 and is projected to reach $45.2 billion by 2026, growing at a CAGR of 44.9% from 2019 to 2026.[1] One of the major challenges in healthcare AI is preserving patient privacy while enabling researchers to access and analyze clinical data. SynGen's platform offers a novel solution by generating synthetic data from large language models, which can be used to train machine learning models without risking the exposure of sensitive patient information. The demand for privacy-enhanced clinical text mining solutions is expected to increase in the coming years, especially with the growing adoption of electronic health records and the need for personalized medicine. Therefore, SynGen's platform has the potential to capture a significant market share and become a major player in the healthcare artificial intelligence industry.

References:

^[1]: Chen, Y., Liu, A., Zhang, L., & Wang, J. (2021). Overview of Clinical Text Mining and Its Applications in Precision Medicine. Frontiers in Genetics, 11, 1-8. doi: 10.3389/fgene.2020.622811

^[2]: Wang, Y., Huang, C., Zhu, W., & Chen, J. (2021). Privacy-Preserving Clinical Text Mining: Current Status and Future Directions. IEEE Journal of Biomedical and Health Informatics, 25(4), 1236-1244. doi: 10.1109/JBHI.2020.3035419

^[3]: Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405. doi: 10.1038/nrg3208

(1.1) Tang, R., Han, X., Jiang, X., & Hu, X. (2023). Does Synthetic Data Generation of LLMs Help Clinical Text Mining?. *arXiv preprint arXiv:2303.04360*.

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2021). Can AI help reduce disparities in general medical and mental health care?. npj Digital Medicine, 4(1), 1-10.
Denny, J. C., Peterson, J. F., & Rosenbloom, S. T. (2013). Challenges in mining electronic health record data: the pediatric experience. Journal of the American Medical Informatics Association, 20(1), 49-57.
MarketsandMarkets. (2020). Clinical data analytics market by type, application, component, delivery, end user - global forecast to 2025. Retrieved from https://www.marketsandmarkets.com/Market-Reports/clinical-data-analytics-market-905.html
Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604.

Zhang, Y., Chen, T., Li, Y., & Wang, S. (2020). Large-scale language model in healthcare: Opportunities, challenges, and strategies. Journal of Healthcare Engineering, 2020.

HealthVerity. (n.d.). Synthetic Data as a Service. HealthVerity. Retrieved March 21, 2022, from https://healthverity.com/synthetic-data-as-a-service/

Linguamatics. (n.d.). Clinical Text Mining. Linguamatics. Retrieved March 21, 2022, from https://www.linguamatics.com/clinical-text-mining

[1]: "Healthcare Artificial Intelligence Market Size, Share & Trends Analysis Report By Offering (Hardware, Software, Services), By Technology (Machine Learning, NLP), By Application (Robot-assisted Surgery, Virtual Nursing Assistant), By End Use, By Region, And Segment Forecasts, 2019 - 2026." Grand View Research, Inc., 2019, https://www.grandviewresearch.com/industry-analysis/healthcare-artificial-intelligence-market.

[1.A] "Clinical Data Analytics Market Size, Share & Trends Analysis Report By Product (Software, Services), By Deployment (On-premise, Cloud-based), By End-use (Payer, Provider), By Region, And Segment Forecasts, 2018 - 2025." Grand View Research, Inc., 2018, https://www.grandviewresearch.com/industry-analysis/clinical-data-analytics-market.

[2.A] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating Multi-label Discrete Patient Records Using Generative Adversarial Networks. arXiv preprint arXiv:1703.06490.

[1.i] Wang, Y., Huang, C., Zhu, W., & Chen, J. (2021). Privacy-Preserving Clinical Text Mining: Current Status and Future Directions. IEEE Journal of Biomedical and Health Informatics, 25(4), 1236-1244. doi: 10.1109/JBHI.2020.3035419

[2.i] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating Multi-label Discrete Patient Records Using Generative Adversarial Networks. arXiv preprint arXiv:1703.06490.

[3.i] Guo, Y., Liu, S., & Wang, H. (2020). Synthetic data generation for healthcare analytics: A review. IEEE Journal of Biomedical and Health Informatics, 24(1), 31-44. doi: 10.1109/JBHI.2018.2844298.