Inquirable Models

Andrew Ghafari
Computer Science

Jay Jhaveri
Computer Science

Vivek Sharma
Data Science

Niraj Yagnik
Computer Science

Abstract

This study explores the integration of Large Language Models (LLMs) with statistical models in healthcare, focusing on their ability to interpret and explain medical risk assessments. Utilizing SHAP values for model understanding, we examined the efficacy of various LLMs, including GPT-4 and its custom variant, GPT-3.5, and LLama2 7B, in providing clear, actionable insights from complex healthcare data. Our methodology involved both qualitative and quantitative evaluations, assessing the models' accuracy in risk prediction and their understandability from both medical and patient perspectives. The surveyed findings reveal significant potential in using LLMs for enhancing the transparency of machine learning models in healthcare, though acknowledging limitations related to prompt sensitivity and the complexity of model algorithms. This research paves the way for future advancements in patient-centric, interpretable AI in the medical field.

1 Introduction

In a rapidly evolving landscape of healthcare, the utilization of statistical models and large language models (LLMs) for risk assessment has become ubiquitous (Meskó, 2023). These models help in predicting outcomes and formulating treatment plans. However, a critical challenge that persists is the interpretability and explainability of these models. The ability to comprehend and trust the inferences drawn by statistical models is crucial, particularly for people who are not extremely well-versed with the medical industry (Kaddour et al., 2023). Everyone in the computer science industry is now exceptionally interested in doing more extensive research about the explainability of these models, and we decided to take a small share of that by working on the project title "Inquirable Model". And since we also had a member of our group with previous experience in the exAI field, this felt like a perfect fit for us. In this project, we decided to focus on enhancing the transparency of machine learning models, allowing patients to query risk assessments using simple natural language. We had a few research questions that were key driver in our research, and we have tried to use different LLMs to compare their efficiency in answering these questions.
These research questions are:

- Can LLMs be used for explaining Medical Risk Models
- Can LLMs be used for potentially reducing the Risk Scores
- What is the best LM we can use for this purpose?

By figuring out answers or partial answers to these questions we aim to bridge the gap between complex algorithmic decisions that sit behind these black box model and true patient understanding. This could ultimately revolutionize how we interact with, understand, and ultimately, trust the technology that is increasingly at the heart of healthcare decision-making.

2 Literature Survey

For our research, and in our pursuit of personalized medicine for everybody, we have decided to incorporate SHAP (Lundberg and Lee, 2017) library and shapley values in our model understanding. This library helps quantifying an individual contribution score for each feature used to determine a predictive model's output. By exhaustively iterating over all possible feature permutations, Shapley values measure the marginal impact on the prediction when a feature is included/excluded within/from a subset of features which helps us in determining the most important and impactful ones. Through (Slack et al., 2023) we then analysed how the model

can be integrated with a large language model using the shapley values for explainability. The paper proposes "TalkToModel," an interactive dialogue system designed to explain machine learning (ML) models through natural language conversations. As ML models have become increasingly complex and challenging to understand, there is a need for effective explainability methods. TalkToModel addresses this need by providing a user-friendly way for practitioners to interact with and understand ML models. The paper,(Bienefeld et al., 2023), explores the differences in how developers and clinicians perceive explainable artificial intelligence (XAI) in healthcare, identifying conflicting goals and perspectives. It proposes design solutions to bridge these gaps, enhancing the effectiveness of XAI systems in clinical decision support. The work to provide explainability through large language models ,follows from the few shot learning approach discussed the paper (Brown et al., 2022) which introduces GPT-3, an autoregressive language model with 175 billion parameters, which demonstrates a significant improvement in task-agnostic, few-shot performance. Through works based on (Meta AI Research Team, 2023)(OpenAI, 2023a)(OpenAI, 2023b), we further explore a wide range of capable Large Language models for the task of interpretability and explainability through natural lanaguage.

3 Datasets and Risk Models

We decided to try with two predictive risk models to try and get a broader and more comprehensive idea about the efficacy of different Language Models in different settings. We tried also to provide the best starting position for the LLMs by curating two high performing risk models, and for that we had to do some preprocessing, and hyperparameter tuning for both models that we are going to discuss more of in the paragraph below. We dived into healthcare models that utilizes two datasets: the Diabetes Dataset (Smith et al., 1988) from the National Institute of Diabetes and Digestive and Kidney Diseases and another dataset for Heart Failure Prediction (Chicco and Jurman, 2020) which is a dataset that encompasses 12 distinct features to anticipate heart failure mortality.

We started with dataset acquisition, followed by data preprocessing to ensure quality and compatibility for model training, and we implemented some feature selection techniques. We also trained different Machine Learning models and we also did hyperparameter tuning to refine model performance for the reason stated above. We also had an evaluation stage to assess the efficacy of the model using metrics like accuracy, F1 score, precision, and recall and choosing our optimal one.Accuracy, precision, recall, and the F1 score are chosen because they collectively provide a comprehensive view of a model's performance, indicating not just overall correctness but also how well the model identifies relevant instances and minimizes false positives and negatives.

In the applied models, the Diabetes Dataset found its best match with a Light Gradient Boosting Machine (Light GBM), which had the best results out of all the other models we tried, scoring around 80% in the metrics we use:an accuracy of 0.870, precision of 0.830, recall of 0.800, and an F1 score of 0.814 highlighting its suitability for diabetes prediction. This is further validated by an impressive ROC Curve, showcasing an AUC of 0.94, indicating excellent model performance in distinguishing between the two classes. The feature importance for the Diabetes dataset, as depicted by the SHAP Beeswarm Plot 1c, highlights insulin and glucose levels as significant predictors, reflecting their biological relevance to the disease's progression.
Conversely, for the Heart Failure Prediction dataset the most promising results were found with a Random Forest model, displaying a balanced performance across all metrics, with accuracy and F1 score both above 80%, with an accuracy of 0.848 and an F1 score of 0.861. Particularly impressive was its recall of 0.939, indicating a high sensitivity in predicting heart failure events. Its ROC Curve 2b supports these findings, with an AUC of 0.96, signifying exceptional discriminative power. The SHAP Beeswarm Plot 2c indicates ejection fraction and serum creatinine levels as the most influential factors, which aligns with medical understanding of heart failure risk factors. These results suggest a robust predictive capability, making Random Forest an ideal candidate for this dataset.
Due to their respective performances, these models were selected for further research into explainability, more details of the models' performance can be found in table 1

4 Methodology

We now describe the paper's methodology and how the work integrates the risk models with the large

Table 1: Performance Metrics and Model Parameters for Risk Models

Algorithm	Accuracy	Precision	Recall	F1 Score	Model Parameters
Random Forest	0.848	0.794	0.939	0.861	[max_depth=None] [n_estimators=150, criterion='gini', max_depth=10, min_samples_leaf=2, max_features='sqrt']
Light Gradient Boosted	0.870	0.830	0.800	0.814	

language models. The work draws inspiration from (Slack et al., 2023) on LLM integrations and evaluation analysis. This section is divided into three subsections. The first talks in-depth about the various Large Language Models the work will utilize, the second talks about which specific language model the report will target, and the subsection on prompt engineering dives deep into what the models are exactly fed with.

4.1 Risk Model - LLM Integration

The work uses Shapley values from the risk models to give the language model more context about the model, its features, and the individual patient (Bienefeld et al., 2023; Srinivasu et al., 2022; Slack et al., 2023). The weights help the model grasp which features are pivotal and how they impact the individual patient's risk score. This informed context allows the language model to generate more personalized and accurate recommendations for potentially modifying health-related behaviors or markers to reduce risk scores.

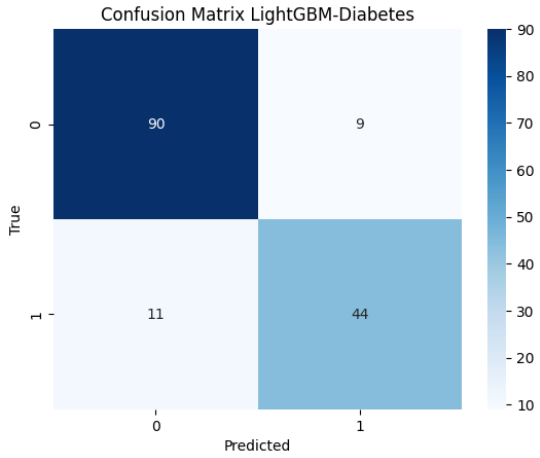
By evaluating the contribution of each feature within the risk model, Shapley values facilitate a granular understanding of how each data point—from biometrics to clinical history—impacts the overall risk score. This attribution of significance to individual predictors enables the LM to generate context-aware health recommendations and personalize its advice to the patient's unique profile. By integrating Shapley's values, the LM becomes proficient in recognizing the interplay of various health determinants in the risk model. It can then suggest modifications in a patient's health parameters that will significantly reduce the risk score. This approach bridges the gap between high-level risk assessment and patient-specific advice, turning complex model outputs into actionable health insights. Consequently, the LM becomes an intermediary, translating statistical predictions into practical lifestyle or treatment changes individuals can understand and act upon. This enhances the usability of risk models in clinical settings.Figure 3 illustrates visually how the integration would work.

4.2 Choice of Language Models

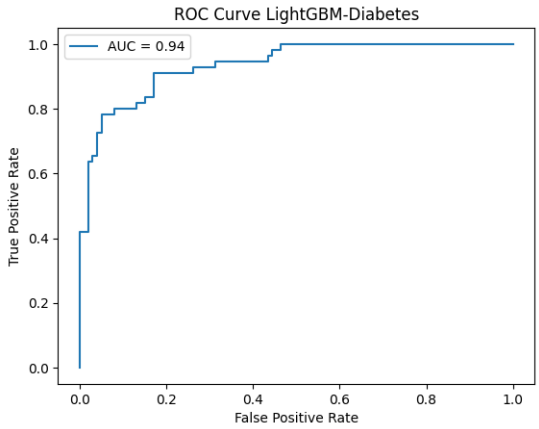
The study explores applying four sophisticated language models to interpret and enhance healthcare risk predictions across two distinct datasets. These models include:

- GPT-4 (OpenAI, 2023a): The latest generational advancement in language models, known for its deep learning and extensive training data, enabling nuanced understanding and generation of human-like text.
- GPT-3.5 (Brown et al., 2022): A slightly earlier version of the generative pre-trained transformer that provides a robust natural language understanding and synthesis framework.
- CustomGPT (OpenAI, 2023b): A tailored language model that allows for customized responses and interpretations based on the unique requirements of the risk model. This variant of ChatGPT can integrate more information via input documents and links to make it more context-aware and educated.
- LLama2 7B (Meta AI Research Team, 2023): A large-scale and open source language model by Meta with a diverse training corpus, designed to generate accurate and contextually relevant text outputs.

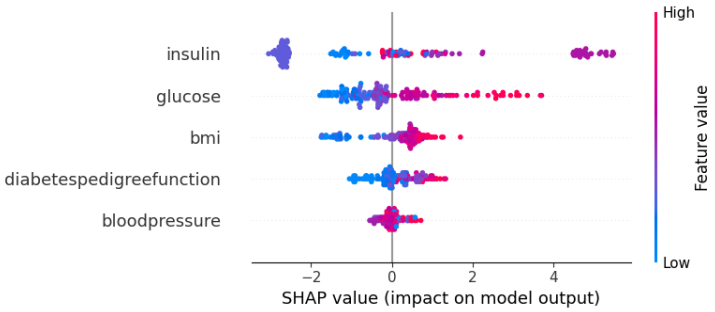
The project aims to translate complex numerical data into comprehensible language to inform patients about their health risks and potential mitigating actions by applying these models to interpret Shapley values extracted from the risk models. The comparison across these models also allows



(a) Confusion Matrix



(b) ROC Curve



(c) SHAP Beeswarm Plot

Figure 1: LightGBM-Diabetes Plots

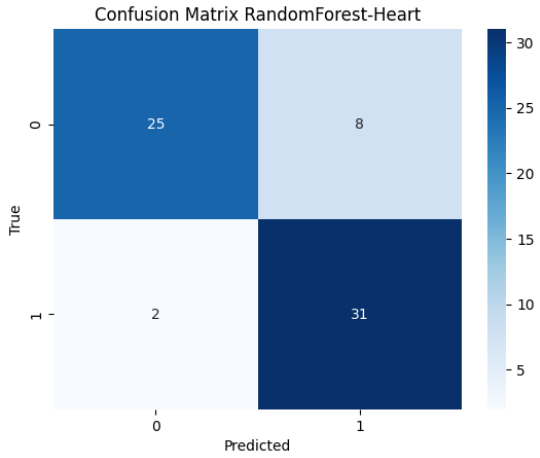
the research to determine which language model is most effective for explaining medical risk models, thereby contributing to medical artificial intelligence and enhancing patient care.

4.3 Prompt Engineering

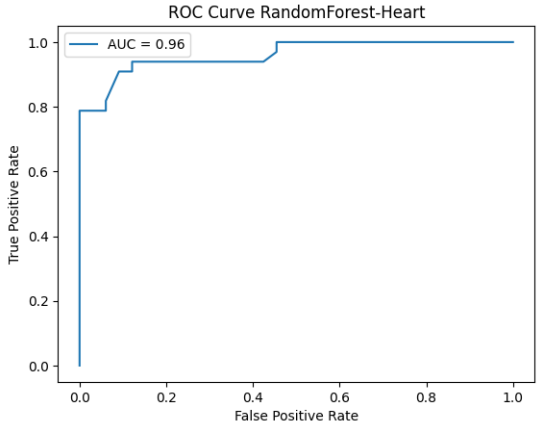
An essential part of the experiments is ensuring the model provides a complete and comprehensive view of the patient and has all the context necessary to answer the input question (Meskó, 2023; Wang et al., 2023). The prompt employed for our experiment is composed of the following elements:

- **Providing Context/Introductory Text:** This indicates the initial input given to the language model, which sets up the scenario or data it needs to consider. This text should be detailed enough to orient the model to the task at hand but concise enough to be easily processed. For our project, we provide the language model information about the project and the expectations for the response.
- **Feature Value:** Feature Value: These are the numerical values associated with the evaluated patient or scenario. In a healthcare setting, these could be vital statistics, lab results, or other quantitative measures necessary to the assessment.
- **Shapley Values:** Shapley values explain the risk model's decisions by indicating the importance of each feature in the decision-making process. Positive values may indicate features that increase the risk score, while negative values suggest features that decrease it.
- **Question:** Upon providing necessary context, the actual question is finally provided to the langauge model. The proposed work primarily focuses on three questins:
  - Why is the risk score so high?
  - What could be done to lower the risk score?
  - How would a change in BMI (or/and, 'glucose,' 'blood pressure,' 'insulin' etc) affect this risk score?

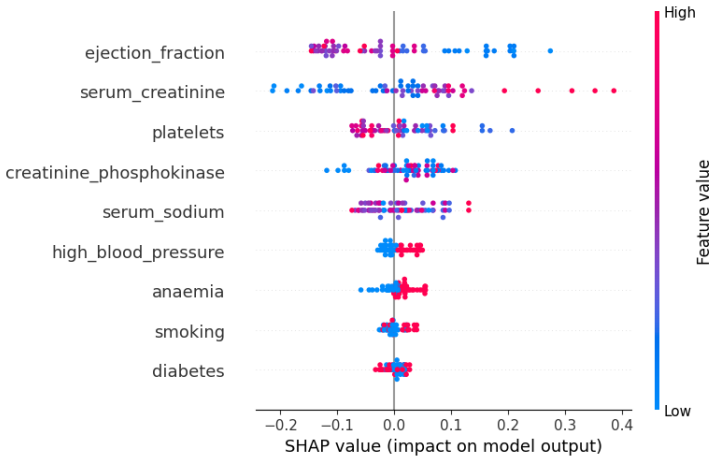
The prompt generator function is made to run for



(a) Confusion Matrix



(b) ROC Curve



(c) SHAP Beeswarm Plot

Figure 2: RandomForest-Heart Plots

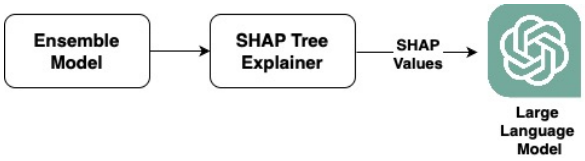


Figure 3: Quantitative Evaluation Process

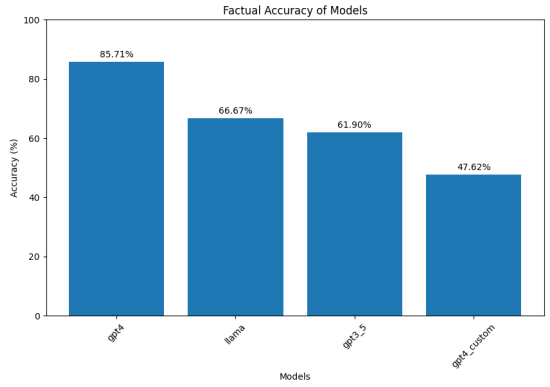
each entry in the test set (equivalent to each patient in the test set).

5 Results

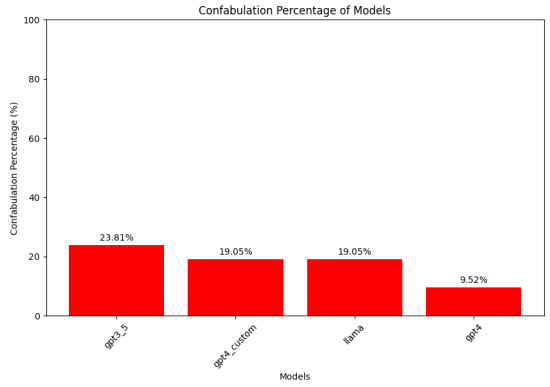
In this section, the proposed work uncovers how well the Large Models perform given the Shapley and feature values corresponding to the patient in the test set along with the corresponding question. The work uses qualitative and quantitative evaluation techniques to get a holistic view of how well models perform. Using the subsequent subsections, we discuss how these evaluations are carried out and how the models perform on them using the respective techniques.

5.1 Qualitative Evaluation

The work employs human evaluation to provide a qualitative measure of the project's success, focusing on the interpretability and applicability of the explanations generated by the language models. The work curates two distinct evaluation forms distributed among doctors and patients to get their feedback on the outputs provided by the LLMs. We had two different interests; we had doctors assess the correctness and relevance of the LLMs' interpretations concerning the current medical standards and their relevance to the case. On the other hand, we also had patients rate the applicability and understandability of the generated explanations, reflecting on how well they could comprehend and apply the suggestions to their daily lives. This dual assessment approach guarantees that the LM-



(a) Factual Accuracy of LLMs



(b) Confabulations % of LLMs

Figure 4: Doctor Survey Results visualized.

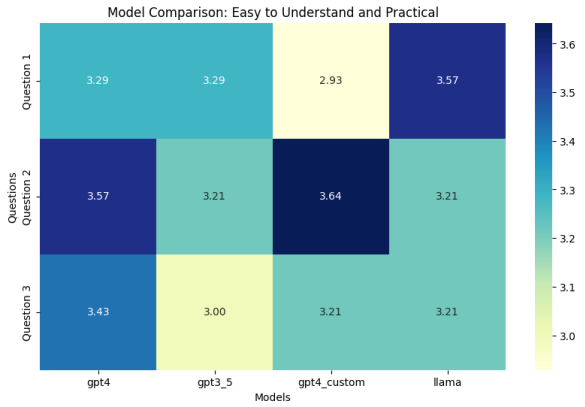


Figure 5: User Responses

generated guidance is not only medically sound but also accessible and practical for patient use, thereby enhancing the overall utility of the interpretability framework.

We managed to evaluate 8 independent doctors to achieve our results for the different large language models we used to gain significant insights, portrayed in Fig 4. GPT-4 triumphed with an exceptional factual accuracy of 85.71%, indicating a highly reliable results. LLAMA follows with a 66.67% accuracy, while GPT-3.5 stands at 61.90%. Notably, the GPT-4 Custom variant shows a lower factual accuracy at 47.62%, which might indicate that while customization can target specific applications, and it may not always provide greater accuracy (Fig 4a). It is worth mentioning that the custom GPT used had natural language prompt that had SHAP documentation, and some background knowledge about some machine learning models and how they operate. Finally it had some example Q&A to follow. More research should be done about how well we can customize GPTs and how to optimize results.

Additionally, the confabulation percentage—indicative of the models' tendency to generate plausible but incorrect information—was assessed as seen in Fig 4b. GPT-3.5 had the highest confabulation rate at 23.81%, followed by both the GPT-4 Custom and LLAMA models at 19.05%. Remarkably, GPT-4 demonstrated a significantly lower confabulation percentage of 9.52%, suggesting that it is the most reliable model in terms of generating medically sound information.

In our examination of the Patient/User Survey responses, our primary objective was to assess the public perception of the diverse Language Model (LLM) responses. We surveyed a target group of 18 independent individuals from different backgrounds. The focus was on determining the clarity and practicality of the suggestions provided by the language models, evaluating their ease of comprehension, and assessing their feasibility for incorporation into one's daily routine within a reasonable timeframe.

The survey results, presented in Figure 5, unveiled an interesting trend. Notably, the LLAMA model consistently received high ratings. Our initial observation suggests that this could be attributed to the concise nature of LLAMA's suggestions, often comprising less than 100 words. In contrast, responses generated by GPT models, while more resourceful, tended to be lengthy.

Furthermore, it is worth highlighting that the GPT 3.5 model exhibited the lowest performance among all the models considered. This observation prompts us to delve deeper into the factors contributing to its comparatively inferior outcome.

5.2 Quantitative Evaluation

The research employs a quantitative evaluation framework to assess the precision of health recommendations provided by large language models about reducing a predicted risk score. The methodology involves modifying individual patient profiles from a test dataset by the LLM's advice. Each piece of advice results in an adjusted patient entry, reflecting changes in specific health metrics, such as body mass index (BMI) and blood pressure (BP). These altered entries are then reprocessed through the risk assessment model to determine whether the LM's recommendations contribute to a lower risk score. The outcome of this validation process serves as a measure of the LM's efficacy in generating meaningful and actionable health guidance. Figure 6 illustrates visually how quantitative evaluation would work.

After getting the suggestions from the model, we would incorporate them into the respective model in inference mode to get the new risk score. Then, we compared and figured out if the risk score increased or decreased and calculated some improvement ratios when possible.

Using the new risk scores, we calculate two metrics, accuracy and % decrease in risk scores. The percentage decrease in probability can be calculated as follows:

$$\% \text{ decrease} = \left(\frac{\text{original prob} - \text{updated prob}}{\text{original prob}} \right) \times 100 \tag{1}$$

The accuracy of the updated risk scores can be determined by the following ratio:

$$\text{accuracy} = \frac{\#\{\text{updated risk scores} < \text{original prob}\}}{\text{total entries in test set}} \tag{2}$$

The quantitative results offered valuable insights into the performance of various Large Language Models (LLMs) in predicting diabetes and heart failure risk. As seen in Table 2 GPT-4 stands out with an impressive 90% accuracy in diabetes prediction and 82% in heart failure prediction, was definitely the best perfoming model overall. Interestingly, the GPT-4 Custom variant shows a reduction in direct accuracy—62% for diabetes and 76% for heart failure—but exhibits a higher percentage decrease in risk scores, 53.52% for diabetes and 51.98% for heart failure, implying a strong potential in patient risk mitigation strategies. GPT-3.5, while the least accurate in direct prediction with 54% for diabetes and 56% for heart failure, still contributed to risk score reduction, but rather modestly.

LLAMA had 70% accuracy in diabetes and 74% in heart failure predictions which can be considered a balanced performance, by had a notable 20.51% reduction in diabetes risk and a minimal 0.81% reduction in heart failure risk. Since we thought that the results of reduction scores in LLAMA were a bit surprising, we dived a bit deeper into the numbers and realized that the model had a very few recommendation which in fact increased the patients' risks of the disease that offsetted the majority of cases where the model actually had really strong recommendations.

6 Limitations and Future Work

The study acknowledges several limitations that we also felt the need to discuss. Primarily, there is a significant reliance on SHAP values to interpret the model's predictions, which, while helpful, can sometimes oversimplify the complexities of the underlying algorithms. Additionally, the models' outputs were found to be sensitive to the prompts provided, indicating a need for careful prompt design to avoid biased or misleading results. Lastly, the necessity for broader human evaluations is emphasized to ensure the models' interpretability and recommendations are robust and generalizable across different populations and clinical scenarios.

The research paper proposes several future directions for further investigation and improvement:

More Health Conditions: The project aims to broaden the scope of its analytical capabilities by incorporating a more comprehensive array of health conditions. This expansion would allow the model to address more diverse clinical scenarios, making it a more comprehensive patient tool. Future efforts will involve integrating datasets that include a variety of chronic and acute conditions, improving the model's ability to provide relevant and precise recommendations for a broader patient population.

Fine-Tuning: The project intends to refine the performance of both T5 and LLAMA models by fine-tuning them on specially curated datasets. The current versions of fine-tuned LLAMA and T5(Raffel et al., 2020) tend to output too technical responses, making it challenging for the patients to understand. For future work, we can focus on curating a dataset that contains nuanced medical information, thus enabling the models to understand better and process complex model works and

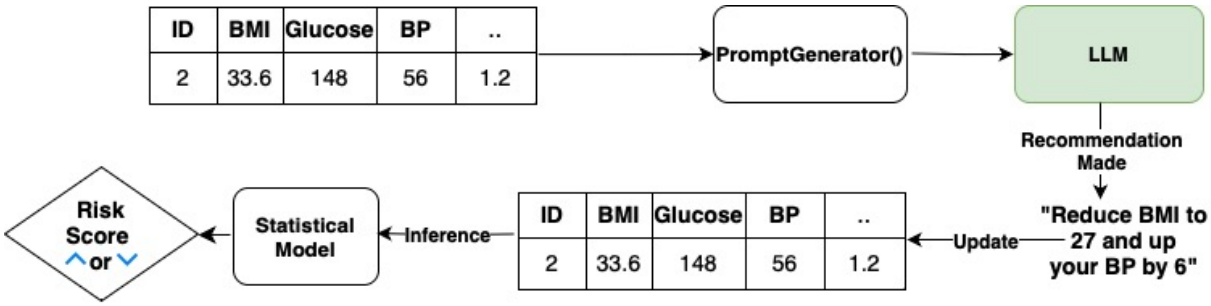


Figure 6: Quantitative Evaluation Process

Model	Diabetes Accuracy	Heart Accuracy	% Decrease Diabetes	% Decrease Heart
GPT4	**90%**	**82%**	43.60%	37.14%
GPT4 Custom	62%	76%	**53.52%**	**51.98%**
GPT 3.5	**54%**	**56%**	**1.13%**	11.75%
LLAMA	70%	74%	20.51%	**0.81%**

Table 2: Model Performance Comparison

translate it for the lay user.

MIMIC Dataset (Johnson et al., 2016): Utilization of the MIMIC (Medical Information Mart for Intensive Care) dataset is planned for more in-depth experiments. The MIMIC dataset, one of the largest publicly available databases of intensive care unit (ICU) patients, provides rich, de-identified health data from a diverse cohort. By applying the models to this dataset, the project can further validate and refine its algorithms, ensuring they are robust and effective across varied clinical settings and patient demographics.

7 Conclusion

The proposed work has successfully demonstrated the effectiveness and reliability of leveraging large language models (LLMs) in healthcare risk assessments. By utilizing the advanced capabilities of LLMs, the project has illuminated the potential to elucidate complex risk models and provide clear, actionable recommendations to mitigate health risks. The language models have shown a significant capacity to interpret intricate Shapley values and patient data, converting them into understandable, personalized health advice.This marks a pivotal step towards integrating AI more deeply into healthcare analytics, with the promise of enhancing patient engagement and facilitating proactive health management.

Code: Github

The code for the proposed work can be found on our Github Repo: Inquirable Models

Ethics Statement

Our research adheres to ethical guidelines and considerations. We have taken steps to ensure privacy and data protection by obtaining necessary permissions and informed consent. We have also made efforts to mitigate biases and evaluate the potential impact on different demographic groups. Transparency and interpretability are prioritized, providing clear explanations of our approach and limitations. We are committed to a safe and inclusive environment and consider the broader societal implications of our work. We welcome feedback and strive for continuous improvement in our ethical practices.

Acknowledgements

We want to express our sincere gratitude to Dr. Michael Hogarth, Dr. Shamim Nemati, and Aaron Boussina for their invaluable guidance and mentorship throughout this project. Their expertise and support have been instrumental in our progress. Without their contributions, our project wouldn't have reached its current stage. We are truly thankful for their valuable insights and unwavering support, which have greatly influenced our journey and

outcomes.

References

Nadine Bienefeld, Jens Michael Boss, Rahel Lüthy, Dominique Brodbeck, Jan Azzati, Mirco Blaser, Jan Willms, and Emanuela Keller. 2023. Solving the explainable ai conundrum by bridging clinicians' needs and developers' goals. *npj Digital Medicine*, 6(1):94.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. 2022. Language models are few-shot learners. In *Proceedings of the National Academy of Sciences*. OpenAI.

Davide Chicco and Giuseppe Jurman. 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):1–16.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jean Kaddour et al. 2023. Challenges and applications of large language models. https://arxiv.org/abs/2307.10169.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638.

B. Meskó. 2023. The impact of multimodal large language models on health care's future. *Journal of Medical Internet Research*.

Meta AI Research Team. 2023. Llama2 7b: A large-scale open source language model. https://ai.facebook.com/blog/llama2-7b/.

OpenAI. 2023a. Gpt-4: Overview and technical details. https://openai.com/blog/gpt-4.

OpenAI. 2023b. Introducing gpts. https://openai.com/blog/introducing-gpts. Accessed: 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5(8):873–883.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. 1988. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association.

Parvathaneni Naga Srinivasu, N Sandhya, Rutvij H Jhaveri, and Roshani Raut. 2022. From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022:1–20.

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.