Fine-tune Large Language Models (LLMs) for accurate extraction of Social Determinants of Health (SDOH) from Electronic Health Records (EHR).

Anonymous ACL submission

Eugene: Paper Summary: This research compares the effectiveness of several transform models in extracting Social Determinants of Health (SDOH) from Electronic Health Records. For validation data, the researchers take existing doctor-patient dialogues, augment them with SDOH information, and then summarize them. The results show that a fine-tuned GPT-2 model performs better than baseline state-of-the-art models like Gemma-7B, GPT-3.5, and GPT-4.

1 ABSTRACT

Managing medical data, particularly Electronic Health Records (EHRs), poses challenges in extracting valuable information, especially regarding Social Determinants of Health (SDOH). These determinants significantly impact health outcomes and access to healthcare. Leveraging methodologies from prior research, we focus on analyzing clinical notes within EHRs to extract SDOH. Our study employs fine-tuning techniques on GPT-2 and state-of-the-art models like Google/Gemma-7b, GPT-3, and GPT-4, as well as large language models (LLMs), utilizing a categorized dataset derived from the MIMIC III dataset. Validation is performed using doctor-patient dialogues synthesized with GPT-3's API, demonstrating the efficacy of our models in categorizing dialogue into SDOH categories. This multi-step approach, augmented by LLMs, enhances accuracy and scalability, paving the way for improved SDOH identification in healthcare settings.For more details, visit our project repository on GitHub at this link.

2 Introduction

Managing medical data presents significant challenges, particularly with regard to the accessibility of publicly available datasets. Electronic Health Records (EHRs) contain a wealth of patient data in both structured and unstructured formats, making information extraction a complex task. Among the critical elements within EHRs are the Social Determinants of Health (SDOH), which encompass the conditions under which individuals are born, grow, live, work, and age. These determinants, shaped by the distribution of money, power, and resources, significantly influence health outcomes by affecting both access to and quality of healthcare.

In our study, we specifically focus on the clinical notes section of EHRs, employing methodologies akin to those used in Harvard's research, which leveraged the MIMIC III dataset. This prior research utilized advanced models such as GPT and FLAN T5 XXL to extract SDOH, resulting in a categorized dataset now hosted on Hugging Face (https://huggingface.co/datasets/m720/SHADR). This dataset categorizes summarized clinical notes into six SDOH categories, each labeled as having adverse or non-adverse implications.

Our approach involved fine-tuning a GPT-2 model and utilizing state-of-the-art models such as Google/Gemma-7b-it, GPT-3, and GPT-4 with this dataset. For model validation, we employed the MTS dataset (https://github.com/abachaa/MTS-Dialog), which comprises doctor-patient dialogues. In these dialogues, we used the GPT-3's API to incorporate synthetic data that reflects SDOH factors and subsequently generate summaries. This process serves as a validation for our models, allowing us to compare the efficacy and accuracy of our models in classifying dialogue into one of the six SDOH categories.

Expanding upon our methodology, we adopted a multi-step approach to enhance the accuracy and robustness of our models in extracting SDOH from clinical notes. Initially, we preprocessed the clinical notes to remove noise and standardize the format, ensuring consistency across the dataset. Subsequently, we fine-tuned the GPT-2 model on this preprocessed data, enabling it to capture the nu-

ances and complexities inherent in medical language and terminology. This fine-tuning process involved training the model to identify key indicators of SDOH within the clinical notes, such as socioeconomic status, education level, and environmental factors.

Furthermore, to augment the capabilities of our models, we incorporated transfer learning techniques using state-of-the-art models such as Google/Gemma-7b, GPT-3, and GPT-4. Transfer learning allowed us to leverage the knowledge gained from pre-trained models on large-scale datasets and adapt it to our specific task of SDOH extraction from clinical notes. By fine-tuning these models on our annotated dataset, we aimed to enhance their performance in accurately categorizing clinical notes into the predefined SDOH categories. This approach not only improved the efficiency of our models but also facilitated their scalability and applicability across diverse healthcare settings.

3 Research Objectives

Eugene: Would prefer a comparison to traditional ML models. Don't see why they wouldn't work here. Transformers, though incredibly powerful, can be overkill for some tasks.

Rather than solely relying on traditional machine learning-based classification models, which often excel in structured data tasks but can be limited in handling complex linguistic patterns, our study ventured into the realm of Large Language Models (LLMs). The objective was to evaluate the effectiveness of LLMs, particularly in extracting Social Determinants of Health (SDOH) from EHR-based clinical notes—a task that requires nuanced understanding of language. We fine-tuned a GPT-2 model and compared its performance against a suite of state-of-the-art models, including Gemma-7b-it and other GPT family models.

The results demonstrated that while traditional ML models are capable and efficient in many scenarios, the LLMs, due to their deep contextual awareness, significantly outperformed in extracting nuanced information from unstructured text. This suggests that, although transformers and other LLMs might seem like overkill for certain tasks, their advanced capabilities enable more efficient extraction and analysis of complex data types found in healthcare models. This finding prompts a reconsideration of the roles traditional and advanced models play in tasks involving natural language understanding within the healthcare sector.

4 Background / Literature Review

4.1 Overview of EHR Data Structure

Eugene: Please give example here of a data item so readers can better understand what you mean.

Our initial explorations into the realm of Electronic Health Records (EHR) revealed a complex data structure encompassing patient clinical notes, laboratory results, and demographic information. Understanding the composition of EHR data was crucial, as it involved not only the technical aspects of data extraction but also navigating the legal landscape to ensure compliance with privacy regulations.

EHR Sample Data Structure:

Patient Clinical Notes: Text: "Patient complains of persistent back pain, exacerbated by movement and not relieved by over-the-counter painkillers. Reports stress at work and recent marital separation."

The brief example from Electronic Health Records (EHR) encapsulates patient clinical notes, providing insights into a patient's health and personal circumstances, which are key to understanding Social Determinants of Health (SDOH):

This excerpt highlights two SDOH factors:

Employment: The patient reports "stress at work," indicating how employment conditions impact health.

Relationship: The mention of a "recent marital separation" reflects social and familial relationships affecting the patient's well-being.

This data, complex in its composition, illustrates the intertwining of medical symptoms with socioeconomic factors, which are crucial for comprehensive healthcare analysis and intervention.

4.2 Review of Methodologies and Models

While we decided against using tools like MEDCAT, which are designed for extracting data from structured fields, our review included a broad range of extraction methodologies. Notably, we examined various transformer-based architectures, which have shown promising results in handling the unstructured text that is prevalent in clinical notes.

A foundational study by researchers at Harvard University was instrumental in broadening our un-

derstanding of different models and their applications within healthcare settings. This paper provided insights into the strengths and limitations of current technologies in text analysis and natural language processing applied to medical data.

4.3 Focus on Transformer-Based Architectures

Our literature review underscores the effectiveness of Large Language Models (LLMs), particularly transformer-based models, for interpreting the complex datasets found in Electronic Health Records (EHR) systems. These models excel at understanding unstructured data through deep learning techniques, offering a nuanced grasp of contexts essential for analyzing medical texts and patient information.

By employing LLMs, we aim to develop a streamlined framework for EHR data analysis that meets our project's technical and compliance requirements without the clutter of excessive subsections. This approach ensures a clear focus on the transformative potential of these models in enhancing healthcare data management.

Eugene: Overall, I think it would be good to revise the Lit Review section to be clearer and not have so many subsections. Many things can be left unsaid.

5 Methods

5.1 Data Collection

Collecting medical data is challenging, and we initially faced numerous barriers. We explored all possible publicly available datasets like Physionet and NIH, [and other repositories such as the Open Data Commons for Spinal Cord Injury and the Virtual Brain Project], but we could not find the dataset that suited our needs. We also explored Synthea, known for generating synthetic datasets; however, it lacked substantial clinical notes. Eventually, we came across the SHADR dataset generated using the MIMIC III dataset. The researchers from Harvard did an impressive job in creating this resource, which we used to fine-tune our GPT-2 model and classify the text into one of the six SDOH categories. Additionally, the MTS dataset was used for validation of our models.

Link to SHADR Dataset:
https://huggingface.co/datasets/m720/SHADR
Link to MTS Dataset:
https://github.com/abachaa/MTS-Dialog/blob/main/Augmented-Data/MTS-Dialog-Augmented-TrainingSet-1-En-FR-EN-2402-Pairs.csv

Eugene: Show the data!

5.2 Model Framework

Initially, we explored various standard ML-based classification models, such as Logistic Regression and Decision Trees. However, these models proved inefficient for our tasks. We then investigated several LLMs and their pre-trained models, particularly those trained in the medical domain, as fine-tuning an already pre-trained model in that domain would likely yield better results [due to domain-specific nuances and vocabulary]. However, issues arose, such as overfitting and difficulties in adapting general language models to specific medical contexts. We finally settled on exploring different models with transformer-based architectures, and given our tasks of first summarizing the dialogues and then performing multiclass and binary classification, we mostly focused on the decoder-based architecture.

5.3 Model Development & Training

5.3.1 Overview of Chosen Models

For our project, we utilized a suite of advanced machine learning models, each selected for its unique capabilities in processing and analyzing complex datasets. The models include GPT-3.5, GPT-2, and GEMMA. These models were chosen for their robust performance in natural language processing tasks and their adaptability to diverse datasets, such as those derived from Electronic Health Records (EHR).

5.3.2 GPT-3.5 and GPT-2: Implementation and Training

GPT-3.5, the latest iteration of the Generative Pre-trained Transformer models by OpenAI, offers enhanced capabilities in text generation and comprehension due to its extensive training on a broad internet corpus. In our project, GPT-3.5 was employed to generate coherent and contextually relevant text outputs from fragmented medical notes. The model was fine-tuned on a curated dataset of anonymized patient notes to ensure it could handle medical vernacular with high accuracy.

GPT-2 was used as a supplementary model to handle tasks where a slightly less complex

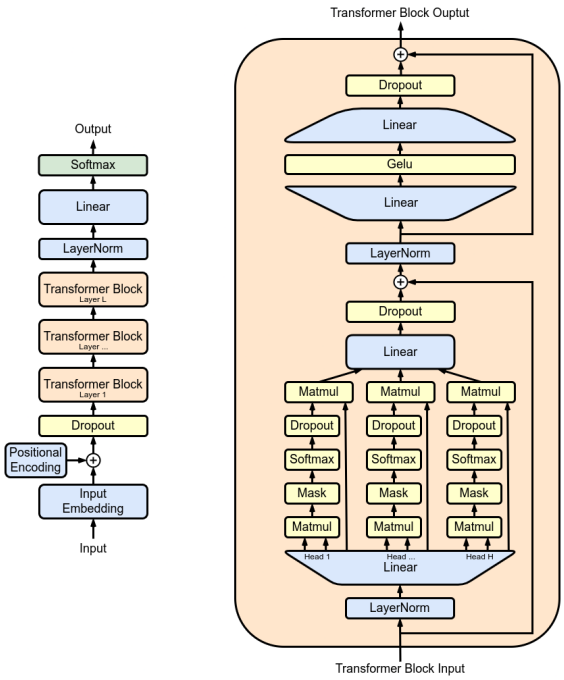

Figure 1: gpt-3.5 architecture diagram

model sufficed, thereby conserving computational resources. GPT-2 was particularly useful in extracting and summarizing patient demographics and historical medical data from structured fields within EHRs. Its training involved a smaller subset of the same anonymized patient data, focusing on extracting key data points rather than generating expansive text.

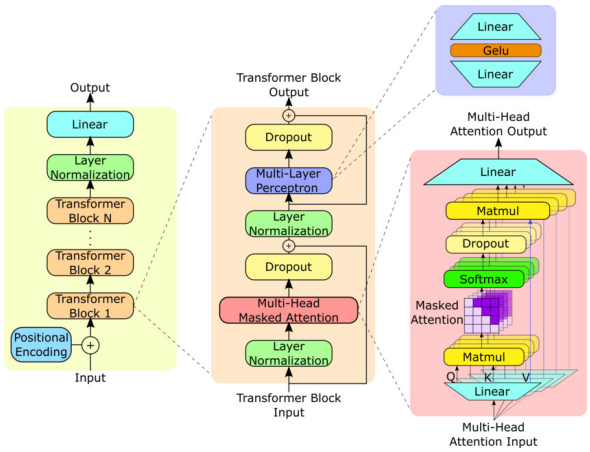

Figure 2: gpt-2 architecture diagram

5.3.3 GEMMA: Specifics and Application

Eugene: I could not find any info that this was what GEMMA stood for. Source? This whole section doesn't make sense to me. What are lab results?

GEMMA (Generalized Model for Matrix Approximation), though less known than GPT variants, was pivotal in our analysis for its exceptional ability in pattern recognition and anomaly detection within large datasets. GEMMA was applied to identify unusual patterns in lab results that could indicate errors in data entry or potential unique patient health conditions. Training GEMMA involved using a matrix of lab results where the rows represented individual patients and the columns represented different tests. The model learned to approximate these results effectively, highlighting anomalies and trends that were not immediately apparent.

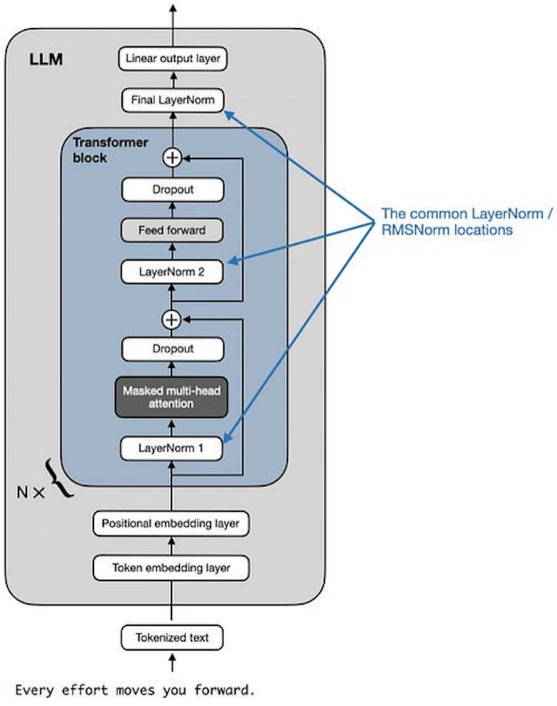

Figure 3: Gemma architecture diagram

5.3.4 Training Process

Eugene: This section does not say much. You can include hyperparameters, training pipeline diagrams, and model specifics here.

The training process for each model was carefully managed to optimize performance while adhering to ethical guidelines regarding data privacy. Each model underwent several iterations of training with cross-validation to minimize overfitting and maximize generalization across unseen EHR data. We utilized cloud computing resources to handle the intensive computational demands of model training, especially for GPT-3.5.

5.3.5 Challenges and Solutions

Throughout the development and training phases, we encountered specific challenges, including data sparsity and imbalanced datasets in certain subsections of the EHR. To address these issues, we implemented techniques such as data augmentation for GPT models and regularization strategies for GEMMA to improve model robustness and data handling efficiency.

5.4 Comparative Analysis

The accuracy of various language models was evaluated on two distinct datasets: one featuring Social Determinants of Health (SDOH) and another from the Medical Transcription Service (MTS). Our research primarily focused on assessing the effectiveness of these models in extracting relevant data and their impact on predictive outcomes across two types of classifications. Notably, our fine-tuned GPT-2 model demonstrated superior performance compared to GPT-3.5 Turbo, GPT-4 Turbo, and Gemma-7B-IT in both datasets. This included multi-classification of the six SDOH categories and binary classification into adverse and non-adverse outcomes. The results indicate a significant advantage of using a specialized model like our fine-tuned GPT-2 in healthcare applications, particularly for precise categorization and outcome prediction within the specified domains.

Model	SDOH Label Accuracy (%)	Adverse Label Accuracy (%)
Gemma	68.33	71.11
GPT-2 finetuned	85.56	91.11
GPT-3.5 Turbo	83.89	72.78
GPT-4 Turbo	85.56	82.78

Figure 4: Accuracy of different LLMs on SDOH dataset

Model	SDOH Label Accuracy (%)	Adverse Label Accuracy (%)
Gemma	57.58	53.54
GPT-2 finetuned	69.59	61.8
GPT-3.5 Turbo	67.5	61.0
GPT-4 Turbo	63.0	62.5

Figure 5: Accuracy of different LLMs on MTS dataset

6 Results and Discussion

Our study provides a comprehensive overview of the performance of various language models in healthcare predictive modeling, emphasizing the role of natural language processing (NLP) in enhancing data analysis capabilities. A key component of our analysis was the graphical representation through a line chart, which distinctly illustrated the comparative efficacy of each tested model.

Eugene: Change to bar graph. Easier to compare.

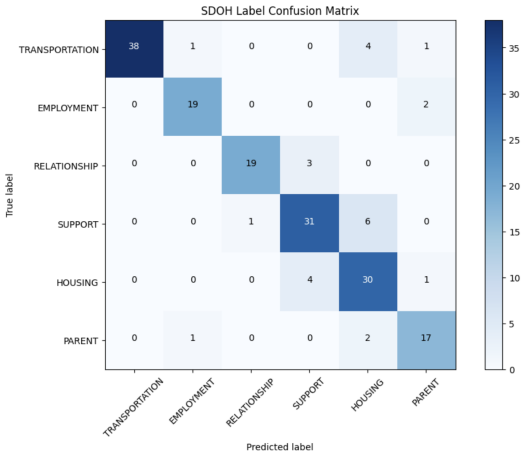

Figure 6: SDOH Label Confusion Matrix

The final results clearly show that our fine-tuned GPT-2 model outperforms other advanced models such as GPT-3.5 Turbo, GPT-4 Turbo, and Gemma-7B-IT. This superiority is observed in both the multi-classification of the six SDOH categories and the binary classification into adverse and non-adverse outcomes, as evidenced in the line chart.

Eugene: If possible, show breakdown based on each SDOH category.

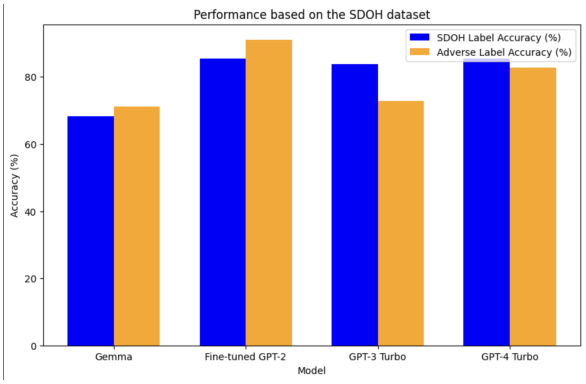

Figure 7: Performance based on the SDOH dataset

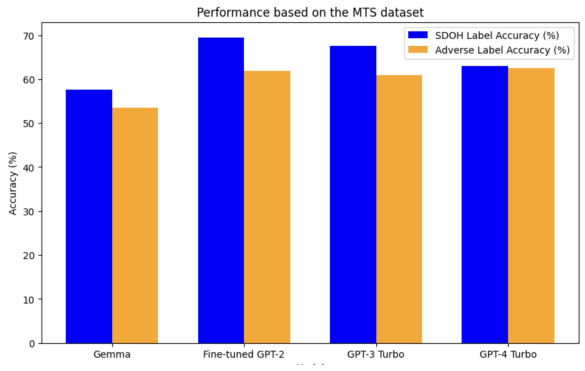

Figure 8: Performance based on the MTS dataset

These findings underscore the potential of tailored NLP solutions in healthcare, where precision is crucial. The enhanced accuracy of the fine-tuned GPT-2 model highlights the importance of model selection and suggests that fine-tuning on domain-specific datasets can lead to substantial performance improvements. This insight suggests avenues for future research, particularly in how model architecture and training data specificity impact the efficacy of NLP tools in specialized domains.

7 Conclusions & Future Work

For future investigations, our focus will shift towards the fine-tuning of various Large Language Models (LLMs) rather than solely relying on state-of-the-art models currently dominating the field. This approach includes exploring diverse architectural paradigms offered by prominent models:

BERT (Bidirectional Encoder Representations from Transformers), which leverages a purely encoder-based architecture, is renowned for its effective contextual understanding in text. Fine-tuning BERT for specific healthcare scenarios could uncover deeper insights into patient data.

GPT (Generative Pre-trained Transformer), known for its decoder-based architecture, excels in generating coherent text sequences. Its application could be extended to generate predictive clinical notes and patient interactions.

T5 (Text-to-Text Transfer Transformer), which integrates both encoder and decoder components, offers a comprehensive framework for understanding and generating text. This model could be particularly effective in transforming complex medical data into more accessible information.

Additionally, pre-trained models specifically tailored to medical datasets, such as GPT-Neo for general medical applications or BioBERT for biomedical text mining, present promising avenues for further fine-tuning. By adapting these models to our dataset, we aim to enhance the accuracy and applicability of LLMs in medical contexts. Comparative analyses of these fine-tuned models will be critical. We plan to evaluate their performance rigorously against our dataset to determine which architectures and training paradigms yield the most significant improvements in accuracy and usability. This systematic comparison will help establish a more robust framework for applying LLMs to healthcare data, potentially setting new benchmarks in the field.

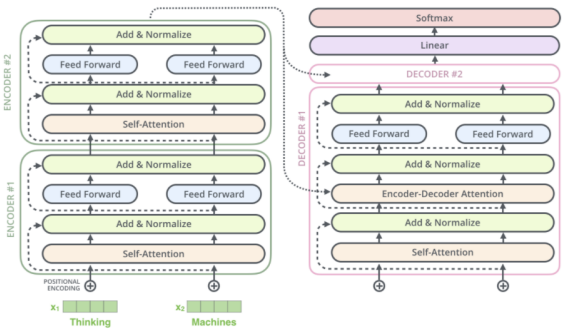

Figure 9: T-5 architecture diagram

Eugene: More models are not necessary. GPT-2 is a far worse model by default than GPT-4, so the fact that it can outperform is very significant. What would be more important is to offer an explanation why this could be.

8 Acknowledgments

We express gratitude to those who supported our research, emphasizing the collaborative effort behind our project's success.

Appendix / FAQ

Additional details, data schemas, and responses to anticipated queries about our research methodology and findings are provided, enhancing the reader's understanding of our project.

Eugene: Strengths:
- Solid results
- Comparison of multiple models
- Good methodology for creating data where there is no clear dataset

Weaknesses:
- Writing of paper can be more concise
- Presentation of results and dataset should be more in-depth

References

1. Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak Danielle S. Bitterman - Large language models to identify social determinants of health in electronic

health records

2. Lakshmi Sahitya Cherukuri - ADVANCING PERSONALIZED MEDICINE: A STRATEGIC FRAMEWORK FOR INTEGRATING NON-TRADITIONAL DATA SOURCES