LLMs in Health Science

Venkata Ramana Reddy Duggempudi and **Pranavi Sriya Vajha**
The State University of New York at Buffalo
vduggemp@buffalo.edu, pvajha@buffalo.edu

Abstract

The use of AI has become crucial in analyzing clinical trial reports (CTRs) to help healthcare professionals to make informed decisions. In this project, an ensemble based Large Language Model which evaluates the truthfulness of the statements in breast cancer CTRs is proposed. The model combines the strength of MedBERT , MedRoBERTa and the Longformer pretrained on medical data to evaluate the reliability of statements in the CTRs. As MedBERT is pre trained in medical data, it has a good understanding of medical context which provides a strong basis for accurate interpretation. MedRoBERTa's optimized attention mechanism ensures that the model understands the medical nuances in the data. Longerformer ability to handle longer texts due to its global attention mechanism helps to understand the detailed CTRs. The ensemble method which combines the logits of these three models using logistic regression is used to predict whether statements entail or contradict the information in the CTRs with high consistency and faithfulness. The development of this LLM increases the trustworthiness of AI in medical decision making by providing a reliable and precise analysis of complex medical data to support health professionals.

1 Introduction

CTRs are comprehensive documents that contain the details and findings of a clinical trial to ensure the safety and efficacy of a medical intervention such as a drug, treatment strategy, etc. The increase in the number of clinical trials and the growth in the clinical trial data made it difficult for medical professionals to understand them. It is difficult for an individual to go through these enormous amounts of text, and understand complex data, putting together different findings to draw useful conclusions is time-consuming. To help medical professionals' researchers are using Artificial Intelligence to analyze summaries of reports, extract key information, and focus on key aspects such as eligibility criteria, treatment specifics, trial outcomes, and observed adverse effects.

The project's main aim is to develop an LLM (Large Language Model) to evaluate the truthfulness of the statements in the CTR, i.e., to predict whether the statement entails or contradicts the information in the breast cancer CTR. The main problem of using AI in this project is to ensure AI can consistently and faithfully comprehend the data in the reports. The AI should understand the minute details and variations in the data which is essential in healthcare.

To overcome these challenges, the project proposes an ensemble-based Large Language model. This model combines the logits of MedBERT , MedRoBERTa and the Longformer pretrained on medical data using Logistic regression. These three models contribute uniquely to the ensemble model. It uses MedBERT's understanding of the medical contexts, MedRoBERTa's optimized attention mechanism and longformer ability to process longer texts due to its global attention mechanism. This ensemble-based model evaluates whether statements in the CTRs entail or contradict the information in the CTRs. The development of this LLMs helps medical professionals to make more faithful and accurate decisions.

2 Literature Review

2.1 MedBERT

The paper from Nature Digital Medicine gives the details about the use of MedBERT. It is pretrained on large-scale biomedical corpora. It is updated to understand the unique lexicon and syntax used in medical domain, which allows the ability to understand medical data and accurately interpret medical documents. The authors demonstrate that MedBERT effectively improves performance on several benchmark datasets for clinical NLP tasks,

which supports its use in the proposed project to enhance the understanding of medical contexts within CTRs.

2.2 MedRoBERTa

The paper in the arXiv publication gives the details about the use of MedRoBERTa. MedRoBERTa which is based on ROBERTa is trained on medical data. The adaptation is important because medical texts use special language and complex sentence structure. RoBERTa outperforms BERT by removing the NSP component entirely, focusing solely on the MLM task. This adjustment was based on findings that NSP does not significantly contribute to model performance for many downstream tasks. It increased the optimizing trainign variables like batch size and learning rate which enhances the comprehension of textual context. By doing these adjustments MedROBERTa is able to process medical data more precisely.

2.3 Longformer

The paper in the arXiv publication gives the details about the use of Longformer. It uses dynamic Masked Language Model approach and eliminates Next Sentence Prediction in training like ROBERTa. It introduces an attention mechanism that combines a global attention with a sliding window approach, allowing the model to process long texts more efficiently. It is Ideal for medical applications involving lengthy documents such as CTRs, Longformer can provide deep insights by processing entire texts in one go.

3 Model Architecture

3.1 Preprocessing Data

The dataset contains three files: train.json, dev.json and gold practice test.json used for training and evaluation of the model.

It also contains CT json folder which consists of detailed information about the patients in different json files.

In preprocessing by using section id gone through the CT json file and extracted the information about every patient. Concatenated the test as follows and created a new columns as concatenated text.

- If the type is single then extracts only primary id details and concatenates the data as a single line.
- If the type is comparison then extracts both primary and secondary id details and concatenates the data as a single line.

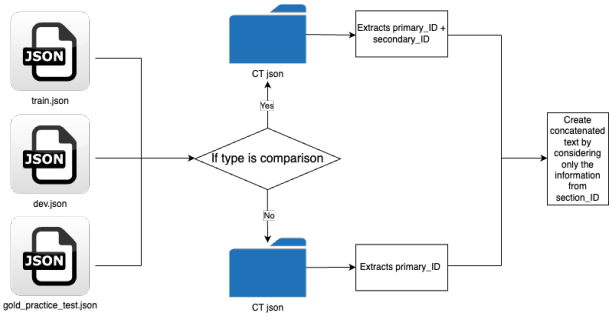

Figure 1: Preprocessing

3.2 Models

3.2.1 MedBERT

MedBERT is a domain-specific variant of BERT, pre-trained for medical data use in applications related to health sciences. It uses the architecture of BERT—a multi-layer transformer network. The MedBERT uses the MLM task during training in which random tokens from the input are masked, and the model is trained to predict these masked tokens. This helps the model understand the context and enhance its language understanding capability that is essential in complicated medical contexts. Similar to the original BERT, MedBERT includes the Next Sentence Prediction (NSP) task, in which the model is trained to predict whether two text segments naturally follow each other. It is helpful in understanding how the flow and relationship of information come in medical documents. It is used to enhance the understanding of medical terminology and context in medical data.

The input that is fed to the MedBERT is tokens. The MedBERT processes these tokens further and generates representative embeddings. The neural network then uses these embeddings, and the model gives the output in the form of logits, which is used to get predictions.

3.2.2 MedRoBERTa

MedRoBERTa is a domain-specific variant of RoBERTa, trained on medical data for its application in health science. It uses RoBERTa architecture. It removes the NSP component completely and focuses on the MLM task only. This was a change based on findings which showed that NSP does not contribute a lot to model performance for

many downstream tasks. RoBERTa, and by extension MedRoBERTa, applies a more dynamic approach towards MLM with a changed masking pattern at each training epoch. This helps the model to learn a more robust understanding of language. It is trained with larger batch sizes and more extensive data, enhancing their performance, especially in domain-specific settings like medicine.

MedRoBERTa takes tokens as input. MedRoBERTa processes the given tokens to represent them as embedding. These embeddings are then fed to the neural network, and the model outputs the logits used to get the predictions.

3.2.3 Longformer

The model used is the Longformer, pre-trained on medical data. Since it is long-range dependencies, Longformer has the MLM pretraining task, similar to BERT and RoBERTa, meaning it is going to be good at comprehending and predicting language in long-range contexts. Longformer also eliminates the need for NSP pre-training tasks, which is just what RoBERTa did when it established that the utility of NSP was minimal in most applications. It introduces an attention mechanism that combines a global attention with a sliding window approach, allowing the model to process long texts more efficiently. This is important to process medical documents of a long range, where the context goes up to the full length of the text.

The input to the Longformer is tokens. The Longformer processes these tokens and produces representative embeddings. These embeddings are then fed to the neural network, and the model outputs the logits, which are then used to obtain the prediction.

3.3 Ensemble Learning

The MedBERT, MedRoBERTa and Longformer methods are combined using ensemble learning method. It combines the strengths and understanding of the models to give accurate, consistent and faithful predictions. The logits of the three models MedBERT, MedRoBERTa and Longformer are concatenated and passed into the logistic regression model to give final output.

- The combined logits ($\mathbf{y}$) from the models are represented as:

$$\mathbf{y} = \begin{bmatrix} y_{\text{BERT}} \\ y_{\text{RoBERTa}} \\ y_{\text{Longformer}} \end{bmatrix}.$$

- The logistic regression function applied on the combined logits is expressed as:

$$\hat{y} = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{y} - \mathbf{b})},$$

where $\mathbf{w}$ is the weight vector and $\mathbf{b}$ is the bias.

The final output, $\hat{y}$, is then used to classify the input as either "Contradiction" or "Entailment".

Logistic regression acts as the ensemble model, effectively leveraging the strengths of each individual model for improved performance.

4 Results

For evaluation we used specified metrics called Faithfulness and Consistency.

1) Macro F1-score: It is calculated from confusion matrix.

The macro F1 score is calculated as follows:

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

where $F1_i$ is the F1 score for the i^{th} class, calculated as:

$$F1_i = 2 \cdot \frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

and N is the number of classes.

2) Faithfulness: It is the measure of the ability of a model to correctly change its prediction when exposed to a semantic altering intervention.

The Faithfulness is calculated as follows:

$$\text{Faithfulness} = \frac{1}{N}\sum_{i=1}^{N} |f(y_i) - f(x_i)|$$

where $x_i \in C : \text{Label}(x_i) \neq \text{Label}(y_i)$, and $f(y_i) = \text{Label}(y_i)$.

3) Consistency: It is the measure of a system to predict the same label for semantically equivalent input.

The Faithfulness is calculated as follows:

$$\text{Consistency} = \frac{1}{N}\sum_{i=1}^{N} 1 - |f(y_i) - f(x_i)|$$

where $x_i \in C : \text{Label}(x_i) = \text{Label}(y_i)$.

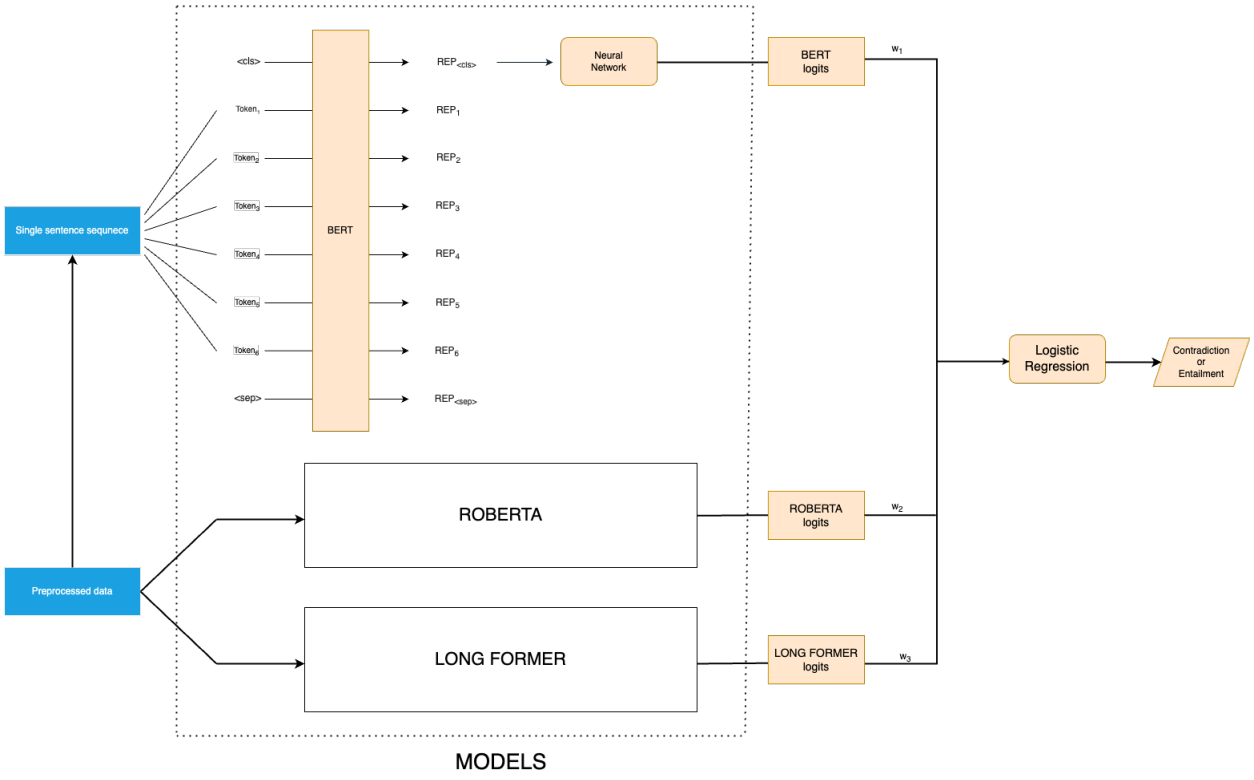

Figure 2: Architecture

We trained the BERT model for 50 epochs and got the training accuracy as 52.18%. When we trained with Med BERT model, Med RoBERTa and Longformer for 100, 90 and 50 epochs respectively. we got an accuracy of 83.35%, 86.41% and 85.41%.

Loss plots for MedBERT, MedRoBERTa and Longformer models while training:

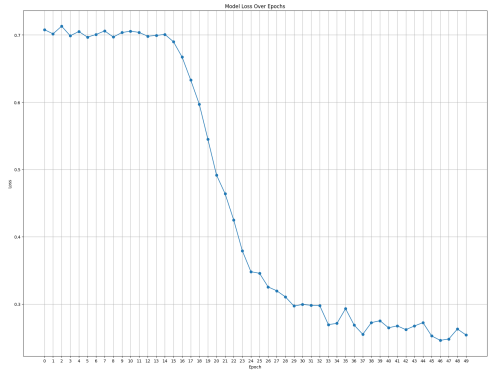

Figure 3: Loss plot for MedBERT model

We validated our model on dev.json and we got the following results for BERT and Med Bert models.

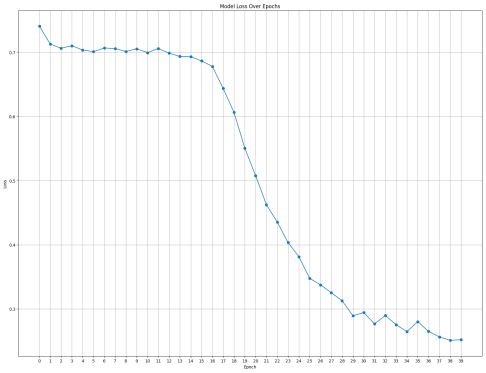

Figure 4: Loss plot for MedRoBERTa model

We used gold practise test data to extract the original statements and their corresponding paraphrased and contradicted statements to calculate faithfulness and accuracy.

This data contains a section called Intervention which has Contradiction, Paraphrase, Text appended. The text appended statements are used for original statements, paraphrased sentences are considered as semantically equivalent sentences which is used to calculate consistency and the contradicted statements are considered as semantically altered sentences which is used to calculate consistency.

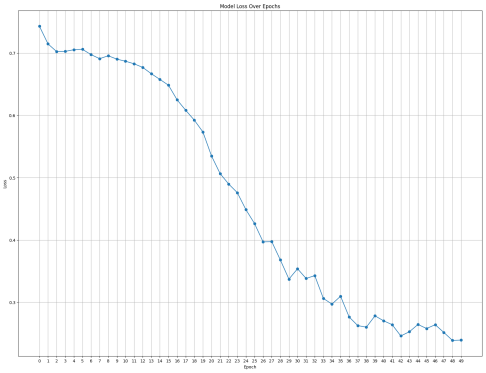

Figure 5: Loss plot for Longformer model

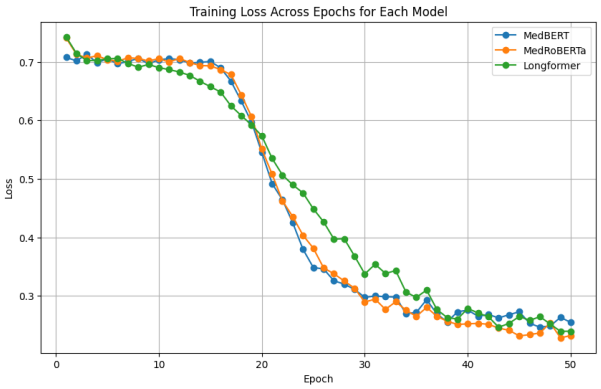

Figure 6: Loss plot comparison for MedBERT, MedRoBERTa and Longformer

Model	Macro F1	Accuracy	Loss
BERT	55	56.2	0.690
Med BERT	53.5	55.8	1.373
Med RoBERTa	53.3	53.8	1.695
Longformer	61.5	61.5	1.060

Table 1: Performance comparison of models.

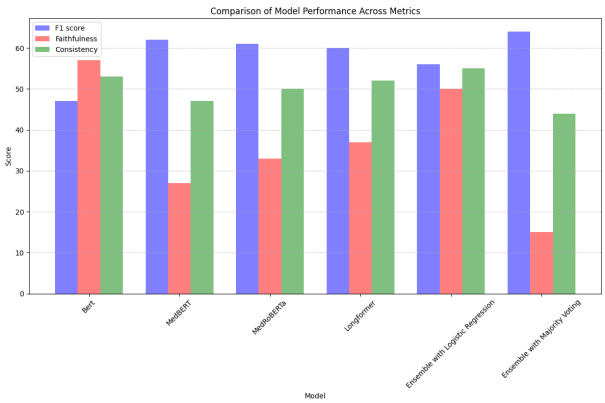

Figure 7: Comparison between models

From figure 6, When the training loss of MedBERT, MedRoBERTa and Longformer is compared the MedBERT has high training loss compared to other models. The training loss for MedRoBERTa and Longformer is similar.

4.1 Comparison between milestone 2 and milestone 3

In milestone 2, we trained only with BERT and MedBERT. The Macro F1-score, faithfulness and consistency for baseline model is 0.62, 0.27 and 0.47 respectively. The baseline results suggest that the model is struggling when semantic modifications are done in the statements. The baseline model is not robust enough to handle variations in the language.

In milestone 3 , we used more optimized models like MedRoBERTa and Longformer which increasesd faithfulness and consistency significantly. The three models are combined using ensembling techniques to increase the understanding of the model and to gain collective intelligence. The ensemble-based model outperforms baseline models with high consistency and faithfulness. This means that the model can handle semantic modifications in the data effectively compared to baseline models.

From figure 7. Consistency, Faithfulness and Macro F1-score is high for ensemble learning model with logistic regression. Even though BERT also has similar results most of the predictions in BERT model are biased to one class. Overall, the ensemble learning based model with logistic regression is reliable compared to other models.

5 Analysis based on results

The ensemble model exhibits good performance across the evaluation metrics for classifying the statements. The **Macro F1-score** for the test results is **56** which is a moderate score for classifying statements but there is a scope of improvement especially considering the critical nature of the medical decision making. The Ensemble model with logistic regression got a score of **50** for **Faithfulness** and **55** for **Consistency**.

- The models which are pretrained on medical data are performing well compared to other models.
- The statements which are simpler are predicted correctly by all three models.
- But as the complexity of the statement increases Longformer is outperforming Med-

Model	F1 score	Faithfulness	Consistency
Bert	47	57	53
MedBERT	62	27	47
MedRoBERTa	61	33	50
Longformer	60	37	52
Ensemble learning with logistic regression	56	50	55
Ensemble learning with majority voting	64	15	44

Table 2: Evaluation metrics of models and test dataset.

BERT and MedRoBERTa.

- The ensemble method trained with logistic regression is performing better when the statements are more complex.
- In this process we combine the understanding of all the models to which increases the collective intelligence to make more accurate predictions.
- The faithfulness and the consistency of the model is highest in the case of ensemble learning. Which is increased by approximately 20% compared to individual models.
- This is because the models can understand the semantic alterations in the text better in the case of Ensemble learning.

Areas of improvements

The model can be improved by using Data Augmentation technique before training the model. For example techniques like synonym replacement, paraphrasing and back translation. It helps the model to learn language variation. This can make the model more robust and can help to improve the performance of the model.

6 Conclusion

The models which are pretrained on Medical data are performing better than base models. As a statement complexity increases Longformer and Ensemble Learning with logistic regression are performing better. After implementing Ensemble learning we got better values for Faithfulness and Consistency. This is because it uses collective intelligence and can understand semantic alterations in the text better than individual models.

To sum up, the ensemble approach with LLM model demonstrates great results in truth evaluation of statements in reports on breast cancer clinical trials. To enhance trustworthiness of artificial intelligence in medical decision-making, the model supplies reliable and precise analysis of complicated medical data.

Contribution

Contributor	Contributions
Ramana Reddy	Literature Review, Preprocessing, BERT, MedRoBERTa, Ensemble Learning with logistic Regression, Code (Preprocessing, models fine tuning and error analysis) Report (Abstract, Literature Review, Model Architecture diagram, Results, Analysis, Improvements and Conclusion) and Git Readme and submission
Pranavi Sriya	Literature Review, Preprocessing, XLnet, MedBert, Longformer, Ensemble Learning with majority voting, Code (Preprocessing, models fine tuning and error analysis), Report(Introduction, Models, Preprocessing Architecture diagram, Analysis, Improvements and Conclusion) and Git Readme

Table 3: Contribution on project

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. 2024. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023.

SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction.

Bram van Es, Leon C. Reteig, Sander C. Tan, Marijn Schraagen, Myrthe M. Hemker, Sebastiaan R. S. Arends, Miguel A. R. Rios, and Saskia Haitjema. 2022. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods.

(Bressem et al., 2024) (Rasmy et al., 2020) (Beltagy et al., 2020) (Devlin et al., 2019) (van Es et al., 2022) (Liu et al., 2019) (Jullien et al., 2023)