LLMs in the Health Sciences

Srija Alla
UBIT: lalla
Buffalo, USA
lalla@buffalo.edu

Vishnu Teja Jampala
UBIT: vjampala
Buffalo, USA
vjampala@buffalo.edu

Abstract

In this report, we introduce a baseline system designed for the Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) task, specifically focusing on Clinical Trial Reports (CTRs) concerning breast cancer treatments. The core of our system is a Sebis Pipeline with a fine-tuned DeBERTa v3 model optimized to evaluate the inference relationships between the reports and statements about crucial treatment aspects. Our methodology involves detailed architectural insights of our model, including its configuration and training processes.

The report thoroughly analyzes the performance of our baseline system, presenting empirical results alongside ablation studies to pinpoint the contributions of various model components. These results help identify our approach's strengths and weaknesses, providing a clear pathway for future enhancements.

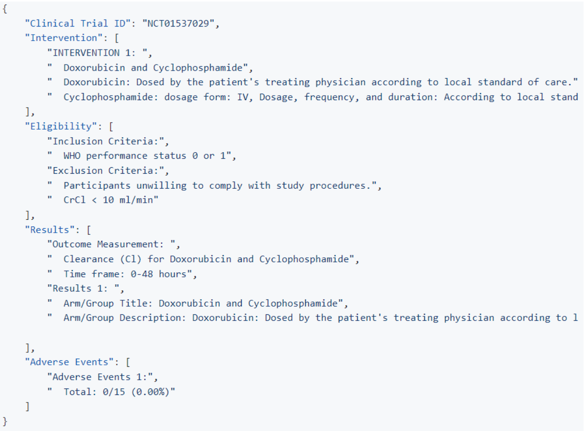
Additionally, we explore potential areas for improvement, such as enhancing the model's ability to handle diverse and complex clinical data, improving interpretability, and expanding training datasets to cover a broader range of inference scenarios. This discussion aims to pave the way for more accurate and robust models in the field of medical natural language processing.

1 Introduction

In this report, the dataset we will work with is based on a collection of breast cancer CTRs (extracted from https://clinicaltrials.gov/ct2/home)(Jullien et al., 2023), statements, explanations, and labels annotated by domain expert annotators.

Each Clinical Trial Report (CTR) consists of 4 sections, as shown in Figure 1:

- **Eligibility criteria** - A set of conditions for patients to be allowed to take part in the clinical trial
- **Intervention** - Information concerning the type, dosage, frequency, and duration of treatments being studied.
- **Results** - Number of participants in the trial, outcome measures, units, and the results.
- **Adverse events** - These are signs and symptoms observed in patients during the clinical trial.



```
{
    "Clinical Trial ID": "NCT01537029",
    "Intervention": [
        "INTERVENTION 1: ",
        "  Doxorubicin and Cyclophosphamide",
        "  Doxorubicin: Dosed by the patient's treating physician according to local standard of care."
        "  Cyclophosphamide: dosage form: IV, Dosage, frequency, and duration: According to local stand
    ],
    "Eligibility": [
        "Inclusion Criteria:",
        "  WHO performance status 0 or 1",
        "Exclusion Criteria:",
        "  Participants unwilling to comply with study procedures.",
        "  CrCl < 10 ml/min"
    ],
    "Results": [
        "Outcome Measurement: ",
        "  Clearance (Cl) for Doxorubicin and Cyclophosphamide",
        "  Time frame: 0-48 hours",
        "Results 1: ",
        "  Arm/Group Title: Doxorubicin and Cyclophosphamide",
        "  Arm/Group Description: Doxorubicin: Dosed by the patient's treating physician according to l

    ],
    "Adverse Events": [
        "Adverse Events 1:",
        "  Total: 0/15 (0.00%)"
    ]
}
```

Figure 1: Clinical Trial Report

The goal is to determine the entailment relation between the CTR premise and the statement. Therefore, the output is either an entailment or contradiction label. The statement can be an entailment or contradiction from a Comparison of 2 CTRs or from a single CTR and one of the sections of CTR (section_id). So in the training data each case can be of type Single or Comparison.

1.1 Reasoning Challenges

There are five distinct challenges that the model needs to overcome to accurately classify. These challenges are:

- **Commonsense Reasoning:** The model must grasp everyday concepts and make judgments on matters that are typically intuitive to humans.

- **Numerical Reasoning:** The model should be adept at performing basic mathematical calculations and understanding relationships such as comparisons, sequences, and magnitudes.
- **Multi-hop Reasoning:** In certain cases, the model is required to synthesize information from multiple sources of evidence (performing several "hops") to arrive at the correct conclusion.
- **Medical Knowledge:** Some claims necessitate a deep understanding of medical terminology and concepts concerning human health, diseases, medications, or treatments.
- **World Knowledge:** This involves the model's ability to process implicit, non-linguistic information about the external world that is embedded in the claims.

In the following sections, we will discuss a few existing architectures to solve the task and its results and analyze the results achieved. Then, we discuss the main model architecture and compare the results with existing models and experiments. We also address the challenges related to numerical reasoning identified by the task organizers. Then, we identify areas for improvement. Qualitative examples from our implementation are provided, and we discuss strategies to improve the performance.

2 Related Work

. The task of Natural Language Inference (NLI) involves determining whether a logical entailment or contradiction exists between two text segments—a premise and a hypothesis. This area of Natural Language Processing (NLP) has been explored since its early days, initially tackled with rule-based methods and linguistic insights, as noted by (MacCartney, 2009) in 2009.

(Zhou et al., 2023) introduces a multi-granularity system for Clinical Trial Report-based textual entailment and evidence retrieval. The model employs a Multi-granularity Inference Network (MGNet), integrating sentence-level and token-level encoding. It also incorporates SciFive, a T5-based model pre-trained on medical texts, to enhance numerical inference capabilities. The system leverages model ensembling and a joint inference module to improve inference stability and consistency.

(Vladika and Matthes, 2023) incorporates a sophisticated multi-task learning architecture for handling natural language inference and evidence retrieval from clinical trial reports on breast cancer treatments. It utilizes a dual-model approach: a pipeline system and a joint system. The pipeline model independently performs evidence retrieval and textual entailment, using extracted evidence sentences as inputs for determining entailment relations. Conversely, the joint system leverages a shared representation to learn both tasks simultaneously, enhancing the interdependence and accuracy of the processes.

This setup allows the joint system to utilize the full context of the clinical trial report, improving its performance on evidence selection and entailment tasks. Both systems are based on a pre-trained DeBERTa v3 model, ensuring robust feature extraction and semantic understanding. The final implementation combines the outputs of both systems in an ensemble model, optimizing prediction accuracy.

But most suffer when it comes to challenges like Multihop Reasoning, Numerical Reasoning, World Knowledge, and Commonsense Reasoning. Sometimes, the model performance can improve, especially in the case of Sebis, if we finetune the DeBERTa v3 model for one of the challenges, for instance numerical reasoning.

Fine-tuning pre-trained language models (LMs) to improve numerical reasoning involves adapting these models to perform calculations and understand numerical contexts more effectively. This process typically utilizes additional training steps that focus specifically on numerical operations and their applications in text.

For instance, (Geva et al., 2020). (2020) demonstrated an approach where they developed a model called GENBERT by injecting numerical reasoning skills into a pre-trained LM. They achieved this by generating large amounts of synthetic numerical and textual data that require mathematical operations. The model was then pre-trained on this data, enabling it to handle numerical reasoning tasks alongside regular language understanding tasks. The fine-tuning on numerical reasoning datasets significantly improved performance on tasks requiring complex numerical manipulations.

In their experiments, GENBERT was shown to improve its F1 score dramatically on the DROP dataset, a benchmark for reading comprehension that involves numerical reasoning. For example, after pre-training on synthetic numerical data, GENBERT's performance on DROP improved from a baseline of 49.3 F1 to 72.3 F1, indicating a significant enhancement in the model's ability to perform numerical reasoning.

3 Model architecture

3.1 Pipeline System

(Vladika and Matthes, 2023) uses a pipeline system to solve the task, and the model's architecture is shown in 2. In this model, we first train the model to select evidence sentences and use these to predict the inference relation.

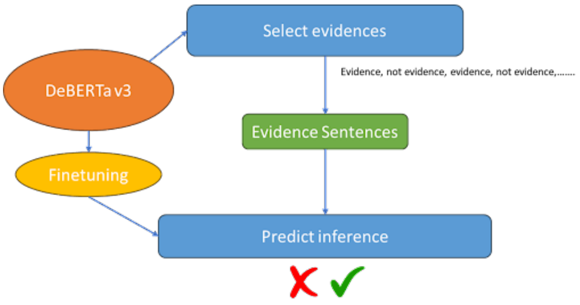


Figure 2: Architecture of Pipeline System

3.1.1 Evidence Selection

In this system, the goal is to train a model that predicts z_i, given a claim c and n sentences $s_1, s_2, s_3, ...s_n$ in the CTR

$$z_i = 1[s_i \text{ is an evidence sentence}] \qquad (1)$$

The candidate sequences are a concatenation of a candidate sentence s_i and a claim c in the form of $a_i = [s_i; SEP; c]$. The dense representations of a_i are obtained as $h_i = BERT(a_i)$. This representation is fed to an MLP and the output is passed through softmax function that assigns the probabilities on whether the candidate sentence is being an evidence sentence or not.

$$p_i, \bar{p_i} = softmax(MLP(h_i)) \qquad (2)$$

The sentences with probabilities > 0.5 are considered evidences. If k final evidences are selected - $e_1, e_2, ...e_k$ are used to input to the next step.

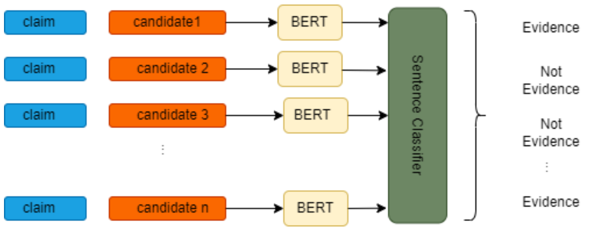


Figure 3: Architecture of Pipeline System for evidence selection

3.1.2 Entailment Task Finetuning

- We tried fine tuning the pre trained DeBERTa-v3-small model on Multi Value NLI dataset
- Dataset for inference having mNLI and DROP examples
- Dataset has premise, hypothesis, label, idx and score
- Label saying if it's an entailment or a contradiction or neutral. We used entailment and contradiction.
- Number of training Examples - 24629 pairs
- Figure 4 shows an example of **DROP** dataset. DROP datasets generally include numerical reasoning and common sense reasoning.

Passage: *Taunton has four art galleries... **Hughes/ Donahue Gallery** founded in 2007, a local community gallery serving local Taunton artists... **Art Euphoric** founded in 2008 has both visual and craft exhibits...*

Q1: How many **years after** founding of **Hughes/ Donahue** was **Art Euphoric** founded?
A1: 1 (number)

Q2: Which gallery was founded **later**, **Hughes/ Donahue** or **Art Euphoric**?
A2: Art Euphoric (span)

Figure 4: Example of Drop dataset Question answering

This Fine-tuned BERT is used to predict inference. Now textual entailment will be a binary classification task that for a given claim c and k, with evidence sentences $e_1, e_2, ...e_k$ we predict if it is entailment or contradiction.

So the claim c is the hypothesis and the concatenation of evidence sentences $e = [e_1, e_2, e_3, ...e_k]$ is the premise. Now, c and e are concatenated as $x = [c; SEP; e]$ and passed to the base language model to get the dense embedding with

$w = BERT(x)$.

$$\hat{y}(c; e) = softmax(MLP(w)) \tag{3}$$

Equation 3 is the probability distribution of each inference label for claim c and evidence e.Doing an $argmax$ on $\hat{y}$ will give us the verdict. The architecture of the pipeline system for evidence selection is shown in Figure 3 and textual entailment in Figure 5

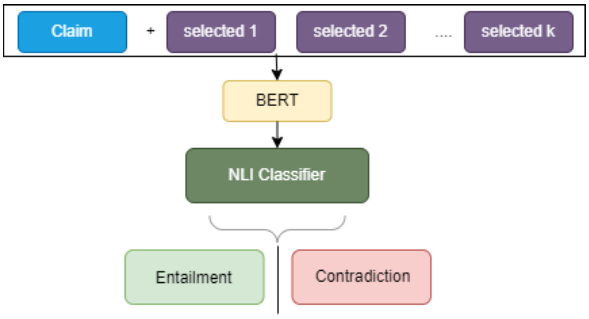


Figure 5: Architecture of Pipeline System for Textual Entailment

4 Experiments and Results

4.1 Analysis of Baseline Models

We started with baseline models (He et al., 2021) domenicrosati/debertav3small-NLI4CT, MoritzLaurer/DeBERTa-v3-small-mnli-fever-docnli-ling-2c and dmis-lab/biobert-base-cased-v1.2. Table 1 shows the scores for Biobert and different versions of DeBERTa v3 and Table 2 shows faithfulness and consistency.

Model	F1	Precision	Recall
BioBERT v1.2	65.93	52.02	90.01
DeBERTa-v3 small	66.07	59.05	75.01
DeBERTa-v3 Moritz	66.66	53.65	88.0

Table 1: F1 score, Precision and Recall scores

Model	Faithfulness	Consistency
BioBERT v1.2	46.87	53.12
DeBERTa-v3 small	41.41	58.59
DeBERTa-v3 Moritz	47.65	52.35

Table 2: Faithfulness and Consistency scores

DeBERTa-v3 Moritz outperforms BioBERT v1.2 and DeBERTa-v3 small in F1 and Recall, suggesting better overall performance and efficiency in retrieving relevant information, though with slightly lower precision than DeBERTa-v3 small. Table 3 compares metrics for Numerical and non-numerical pairs. All models perform better on non-numerical than numerical reasoning, with DeBERTa-v3 Moniz showing the smallest performance drop between categories, indicating better adaptability to numerical data. Figure 6 shows the bar plot of F1 scores for the same.

Model	Numerical	Non Numerical
BioBERT v1.2	63.06	70.58
DeBERTa-v3 small	61.78	72.15
DeBERTa-v3 Moritz	63.15	69.33

Table 3: F1 Score Numerical vs Non Numerical

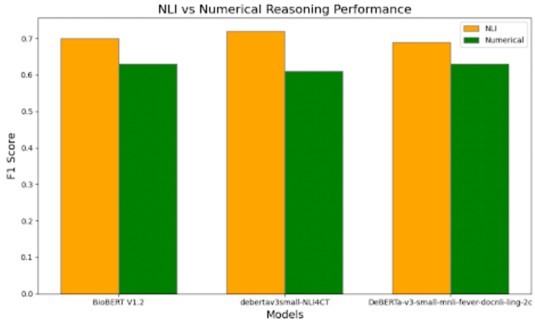


Figure 6: Comparison of baseline models on Numerical vs Non-numerical datasets

Table 4 shows the F1 scores for each section in the development dataset. DeBERTa-v3 Moritz outperforms BioBERT v1.2 and DeBERTa-v3 small across all categories, indicating superior capability in handling diverse clinical trial data segments such as Adverse Events, Results, Eligibility, and Intervention.

Model	Adverse Events	Results	Eligibility	Intervention
BioBERT v1.2	65.71	65.75	68.35	71.1
DeBERTa-v3 small	72.13	66.66	69.33	70.83
DeBERTa-v3 Moritz	68.65	67.53	67.7	73.68

Table 4: F1 Score - Categorised by Sections

4.2 Ablation Results - Sebis

From the results, as DeBERTa v3 Moritz is giving F1 score compared to We tested sebis pipeline system with the above LLMs. Table 5 shows the best F1, Precision, and Recall scores for the different versions of Sebis implementations, with and without finetuning we experimented with for the entailment task and Table 6 shows the Faithfulness and Consistency values for different Sebis versions. As DeBERTa v3 Moritz is giving the best results when used directly, we decided to use this as base language model for Sebis too. Sebis Pipeline used with DeBERTa-v3 Moritz fine-tuned on the larger numerical NLI dataset showed better results compared to other systems.

Model	F1	Precision	Recall
Sebis BioBERTv1.2	62.67	60.0	93.00
Sebis DeBERTa v3 small	63.38	61.11	77.0
Sebis DeBERTa v3 Moritz	64.39	64.2	63.0
Sebis DeBERTa v3 fine-tuned small	68.9	70.21	66.00
Sebis DeBERTa v3 fine-tuned large	72.34	69.56	80.00

Table 5: F1, Precision and Recall scores of Sebis

Figure 7 shows the bar plot of F1 scores with different versions of Sebis. The Sebis DeBERTa v3 fine-tuned large model significantly outshines the others, showcasing the highest scores in all metrics. This indicates that extensive fine-tuning and larger model architecture greatly enhance performance across all metrics, making it the most effective for detailed and complex tasks compared to smaller or less specifically fine-tuned versions.

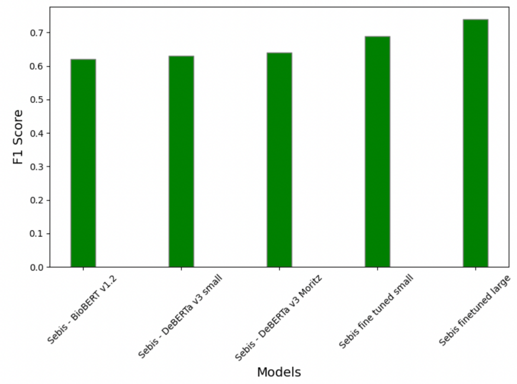


Figure 7: Bar plot with F1 scores of different Sebis models

Model	Faithfulness	Consistency
Sebis BioBERTv1.2	36.0	64.0
Sebis DeBERTa v3 small	36.0	35.48
Sebis DeBERTa v3 Moritz	35.5	64.5
Sebis DeBERTa v3 fine-tuned small	31	69
Sebis DeBERTa v3 fine-tuned large	27.5	72.5

Table 6: F1, Precision and Recall scores of Sebis

Table 7 shows F1 scores achieved by sebis models categorized by sections. The Sebis DeBERTa v3 fine-tuned large model outperforms others in all sections, especially in Intervention, indicating superior adaptation to complex tasks requiring detailed content understanding.

Model	Adverse events	Results	Eligibility	Intervention
Sebis DeBERTa v3 Moritz	63.33	62.19	64.24	63.86
Sebis DeBERTa v3 fine-tuned small	59.23	65.97	74.96	58.30
Sebis DeBERTa v3 fine-tuned large	71.14	62.48	76.77	69.42

Table 7: F1 score - Sebis by Section

Table 8 shows F1 scores for pairs categorised by Type Single and Comparison. As expected, Sebis DeBERTa v3 fine-tuned large significantly outperforms other models in single and comparison tasks, demonstrating enhanced ability to handle complex inferencing across multiple contexts. The Scores for Single are higher than comparison as comparison might require multi-hop reasoning, which is a challenge.

Model	Single	Comparison
Sebis DeBERTa v3 Moritz	62.84	64.75
Sebis DeBERTa v3 fine-tuned small	67.77	59.82
Sebis DeBERTa v3 fine-tuned large	71.42	66.51

Table 8: F1 score - Sebis Single vs Comparison

4.3 Comparison of Sebis with Confusion Matrix

Figure 8, Figure 9 and Figure 10 show the Confusion Matrix of development set for Sebis, Sebis fine-tuned small and sebis fine-tuned large respectively. The confusion matrices show that Sebis DeBERTa v3 fine-tuned large has the highest true positive rate and lowest false negatives, indicating superior predictive accuracy and reliability. Fine-tuning significantly reduces misclassifications, enhancing model sensitivity and precision across tasks.

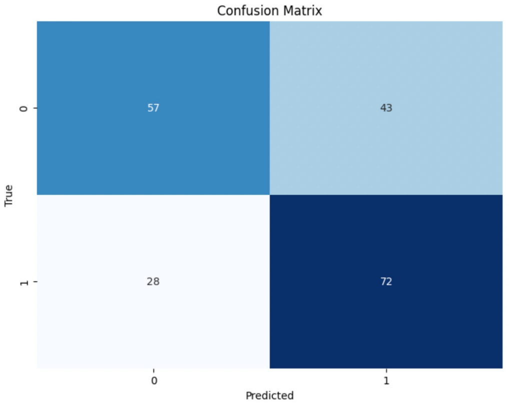


Figure 8: Confusion Matrix of Development set - Sebis

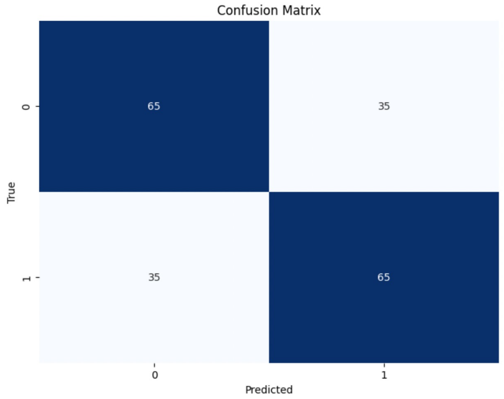


Figure 9: Confusion Matrix of Development set - Sebis Finetuned Small

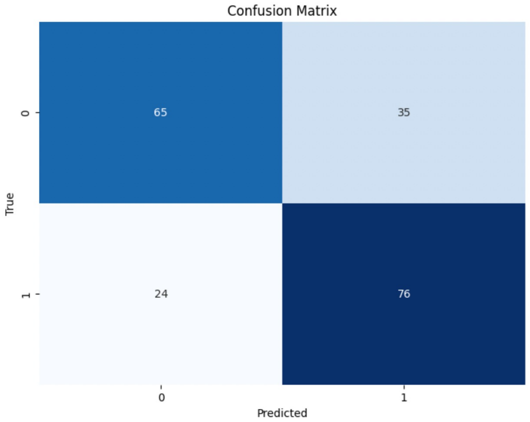


Figure 10: Confusion Matrix of Development set - Sebis Finetuned Large

4.4 Error Analysis

To understand the cases where the model is failing, we made a list of cases from the development set where the model is wrongly predicting the inference relation -

- **Medical Knowledge**
  Claim: Patients must have several visible carcinomas in the stomach and neck areas to be included in the primary trial
  Premise: primary trial: Histologic diagnosis of palpable invasive breast cancer or ductal carcinoma in situ
  Actual Label: Contradiction
  Predicted Label: Entailment
  Predicted Label after finetuning: Entailment

- **Common Sense Reasoning**:
  Claim: only patients with a HER2-positive status can take part in the primary trial
  Premise: primary trial: HER2-positive status (patients who have unknown HER2 status, and for whom determination of HER2 status is not possible, are eligible for this study).
  Actual Label: Contradiction
  Predicted Label: Entailment
  Predicted Label after finetuning: Contradiction

- **Multihop Reasoning**
  Claim: Patients with significantly elevated ejection fraction are excluded from the primary trial, but can still be eligible for the secondary trial if they are 55 years of age or over.
  Premise:Primary trial: Active/symptomatic brain metastases, secondary trial:[....]Cardiac left ventricular function with resting ejection fraction < 50% (below upper limit of normal) [....]
  Actual Label: Contradiction
  Predicted Label: Entailment
  Predicted Label after finetuning: Entailment

- **World Knowledge**
  Claim: The shortest PFS in cohort 1 of the primary trial was 1.4 months below average
  Premise:[...] Baseline, every 6 weeks of study treatment period [...]
  Actual Label: Entailment
  Predicted Label: Contradiction
  Predicted Label after finetuning: Entailment

- **Numerical Reasoning**
  Claim: the shortest PFS in cohort 1 of the primary trial was under 7 months, in cohort 2 the shortest was just over 7 months
  Premise: primary trial: [...] Unit of Measure: months 9.1 (6.8 to 10.8)secondary trial: [...] Unit of Measure: months 7.1 (5.6 to 10.8)
  Actual Label: Contradiction
  Predicted Label: Entailment
  Predicted Label after finetuning: Contradiction

4.4.1 Comparison on Numerical Reasoning

Most models are known to fail in numerical reasoning, so we generated a numerical dataset from the existing dev set to test. Table 9 shows scores for the numerical test cases. Though BioBERT v1.2 is performing almost as well as DeBERTa v3 Moritz, we decided to go with DeBERTa v3 Moritz based on overall performance. However, Sebis with DeBERTa v3 fine-tuned on a large dataset performed better than others in numerical and non-numerical reasoning tasks. Figure 11 compares the F1 scores of Sebis models on the numerical reasoning dataset.

Model	Numerical Reasoning	Non Numerical Reasoning
BioBERT v1.2	63.06	70.58
DeBERTa-v3 small	61.78	72.15
DeBERTa-v3 Moritz	63.15	69.33
Sebis DeBERTa v3	63.52	61.10
Sebis fine-tuned small	64.54	60.46
Sebis fine-tuned large	74.24	68.75

Table 9: Comparison of F1 on all models

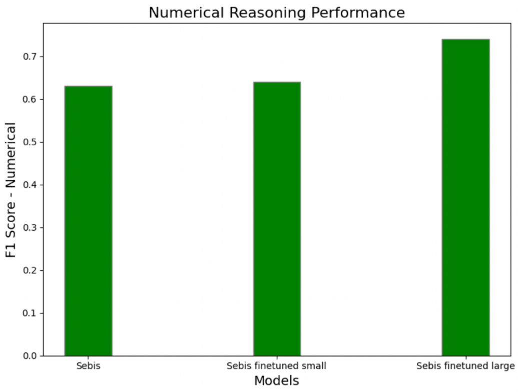


Figure 11: Comparison of Sebis versions on Numerical datasets

If the base language model used in Sebis, fails for numerical reasoning, it is more likely that Sebis will also fail. Table 10 shows the faithfulness and consistency values of models that can be used as base language models.

Model	Faithfulness	Consistency
BioBERT v1.2	45.63	54.37
DeBERTa-v3 small	46.38	53.61
DeBERTa-v3 Moritz	44.48	55.52

Table 10: Comparison of Faithfulness, Consistency

4.4.2 Confusion Matrices

Figure 12, Figure 13 and Figure 14 show the confusion matrices for Sebis models on numerical dataset. .The confusion matrices show improvement from Sebis to Sebis fine-tuned small and large models, with the fine-tuned large model demonstrating the highest true positive rate and lowest false negatives, indicating significantly enhanced prediction accuracy and reliability in numerical datasets.

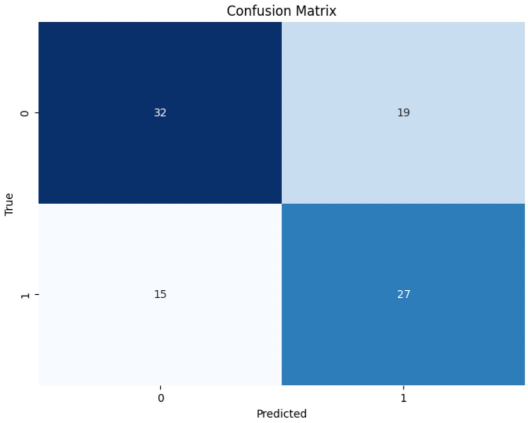


Figure 12: Confusion Matrix of Numerical dataset - Sebis

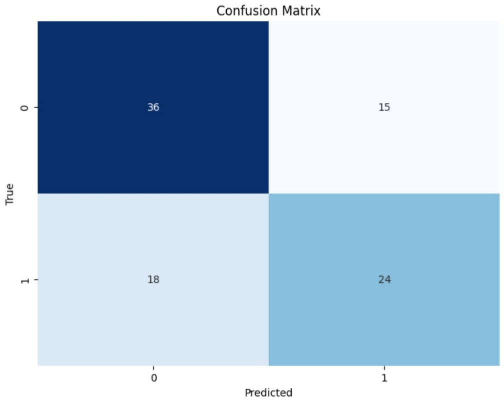


Figure 13: Confusion Matrix of Numerical dataset - Sebis Finetuned Small

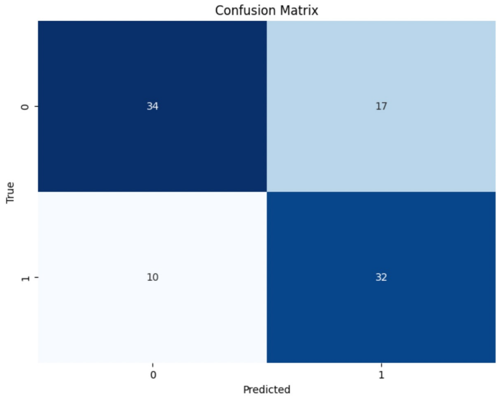


Figure 14: Confusion Matrix of Numerical dataset - Sebis Finetuned Large

4.4.3 Areas of improvement

To enhance our model, we propose several strategies that could substantially boost its performance. First, employing a larger pre-trained language model (LLM) could be beneficial. For instance, switching to the MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli, which possesses 435M parameters, might provide a richer feature extraction capability due to its vast scale and comprehensive pre-training across diverse datasets.

Fine-tuning our model with specialized datasets tailored for medical reasoning can also lead to significant improvements. By adapting the model to recognize and process the nuanced language and complex structures typical in medical texts, we can enhance its applicability and precision in clinical contexts.

These targeted enhancements are aimed at overcoming the current limitations by enriching the model's understanding and responsiveness to the specific demands of medical NLP tasks. By implementing these improvements, we anticipate a marked advancement in the system's overall efficacy, making it a more robust tool for clinical decision support.

5 Conclusion

This paper presents our approach to SemEval-2023 Task 7, which focuses on natural language inference and evidence retrieval from clinical trial reports. We explain the importance of the task, review relevant literature, offer detailed descriptions of the systems we developed, report our findings, evaluate the effectiveness of our models, and explore various challenges encountered during our research.

We used the Sebis pipeline model and enhanced it by finetuning the DeBERTa v3 model, on dataset containing examples of reasoning especially numerical reasoning and commonsense reasoning - MULTI_VALUE_mnli_drop_aux_have, using DeBERTa-v3-small-mnli-fever-docnli-ling-2c. The hyperparameters that gave the best performance were learning rate - 2e-06 with optimizer AdamW and weight decay - 0.06, warmup ratio - 0.1 and batch size = 8. We achieved F1 score of 72.34 and faithfulness and consistency of 27.5 and 72.5, respectively. This outcome demonstrates that fine-tuning has substantially boosted the model's performance. By tailoring the DeBERTa v3 model to the specific characteristics and complexities of a larger and more varied dataset, its ability to accurately predict and analyze data has markedly improved.

Future system enhancements could address challenges in commonsense, numerical, and multi-hop reasoning. Additionally, incorporating more medical and general world knowledge could further improve its capabilities.

6 Contribution

Team Member	Contribution
Srija	Baseline models - BioBERT, DeBERTa v3 small Sebis - Joint System and debugging Pipeline System Sebis joint with different LLMS Experiments - Sebis Pipeline fine-tuning with Large and Hyperparameter tuning Sebis results on Numerical, Non Numerical, Categorised by Sections, Single VS comparison Error Analysis
Vishnu Teja	Baseline Models - DeBERTa v3 Moritz Sebis - Implementing Pipeline System Sebis Pipeline with different LLMS Experiments - Sebis Pipeline Sebis - fine-tuning with Small and Hyperparameter Tuning Extracting Numerical, Non-numerical, Single vs Comparison and Section datasets Baseline model results on Numerical Non Numerical, Categorised by Sections, Single VS comparison

Table 11: Contribution of Teammates

References

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Nli4ct: Multi-evidence natural language inference for clinical trial reports.

Bill MacCartney. 2009. Natural language inference.

Juraj Vladika and Florian Matthes. 2023. Sebis at semeval-2023 task 7: A joint system for natural language inference and evidence retrieval from clinical trial reports.

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. THiFLY research at SemEval-2023 task 7: A multi-granularity system for CTR-based textual entailment and evidence retrieval. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1681–1690, Toronto, Canada. Association for Computational Linguistics.