Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT)

Lekai Chen
University of Colorado, Boulder
lekai.chen@colorado.edu

Meghana Vasanth Shettigar
University of Colorado, Boulder
Meghana.VasanthShettigar@colorado.edu

Shubhankar Goje
University of Colorado, Boulder
shubhankar.goje@colorado.edu

Abstract

This task is based on a collection of breast cancer CTRs, statements, explanations, and labels annotated by domain expert annotators, as outlined in the task overview paper [source]. We aimed to address the challenges of biomedical natural language inference and improve the F1 score using advanced NLP models. Our system, evaluated against the provided dataset, demonstrates competitive performance compared to baseline models, with notable F1 scores. The report also includes an analysis of various models, including GPT3.5 Turbo, llama-2-dev, PubMedBert, TF-IDF Entailment prediction baseline, and BM25 Evidence Retrieval baseline.

1 Introduction

The SemEval-2024 Task 2, "Safe Biomedical Natural Language Inference for Clinical Trials," is a pivotal challenge in natural language processing, focusing on elucidating the entailment relationship within biomedical statements, thus contributing to the advancement of clinical trials. This task revolves around textual entailment using a dataset derived from breast cancer Clinical Trial Reports (CTRs) sourced from [link]. The CTRs have been condensed into four distinct sections: Eligibility criteria, Intervention details, Results metrics, and Adverse events. Annotated statements, averaging 19.5 tokens, articulate claims about information within these sections, and the goal is to ascertain the inference relation (entailment or contradiction) between CTR-statement pairs, with test set and development set interventions designed to ensure a fair and robust competition. Technical details of these interventions will be disclosed post-evaluation, fostering transparency and encouraging resilient approaches in addressing the task complexities.

The language under consideration encompasses the specialized domain of biomedical sciences. The task's significance lies in its potential to refine information extraction processes crucial for biomedical research and clinical decision-making. This task, as outlined in the task overview paper, addresses the critical need for ensuring the accuracy and safety of natural language understanding within the biomedical context, thereby augmenting the broader field of clinical research.

1.1 Main Strategy:

Our system employs a sophisticated neural network-based architecture, specifically tailored to the biomedical domain, for addressing the entailment prediction task. This strategy leverages advanced deep learning techniques, incorporating pre-trained embeddings and attentive mechanisms to capture intricate relationships within biomedical text. The architecture is meticulously designed to navigate the challenges unique to the biomedical domain, ensuring nuanced representation and inference capabilities.

1.2 Discoveries and System Evaluation:

Participation in this task has yielded insightful revelations regarding the complexities inherent in biomedical natural language inference. Our system achieved commendable quantitative results, notably in terms of F1 score, precision, and recall. Comparative analysis against other state-of-the-art models reveals the competitiveness of our approach. However, our system encountered challenges in accurately capturing subtle entailment relationships, particularly in instances where contextual nuances heavily influenced the inference process. This underscores the inherent complexity of biomedical language and highlights potential avenues for future research to enhance the system's robustness. We are committed to fostering research collaboration and transparency. To facilitate the reproducibility of our work, we have released the complete

source code for our coursework [github].

2 Supervised Finetune LLAMA-2 in Text Generation for Classification

2.1 Why Llama 2?

The Llama 2 release represents a significant leap in the field of Large Language Models (LLMs), offering a range of models from 7B to 70B parameters. These models are not only larger in scale but also exhibit substantial improvements in terms of training data (40% more tokens) and context length (up to 4k tokens). The introduction of grouped-query attention mechanisms further enhances inference speed, making Llama 2 an ideal choice for demanding NLP tasks.

One of the standout features of Llama 2 is its fine-tuned versions, particularly Llama 2-Chat, which have been optimized for dialogue applications using Reinforcement Learning from Human Feedback (RLHF). This approach has led to significant improvements in helpfulness and safety benchmarks, positioning Llama 2-Chat models on par with industry leaders like ChatGPT. The detailed insights into this advancement can be found in the Llama 2 release paper.

2.2 Llama-2-7b-chat-hf

We select the Llama-2-7b-chat-hf as our base model. The Llama-2-7b-chat-hf model is a part of the Llama 2 family, which represents a significant advancement in the field of Large Language Models (LLMs). This specific model, with its 7 billion parameters, is designed to strike a balance between computational efficiency and the ability to handle complex natural language processing tasks. Here's a deeper look into its features and capabilities:

- **7 Billion Parameters**: 7B parameters make it significantly powerful, yet more manageable in terms of computational resources compared to its larger counterparts. While powerful, the 7 billion parameter model does not demand as extensive computational resources as larger models, making it more accessible for research and practical applications.

- **Pretrained on Diverse Data**: The model is pretrained on a vast and diverse dataset, enabling it to understand and generate a wide range of text types and styles.

- **Fine-Tuned for Dialogue Applications**: Specifically optimized for chat and dialogue applications, the model has undergone fine-tuning to improve its conversational abilities. This includes better understanding of context, generating coherent and contextually relevant responses, and maintaining the flow of conversation.

- **Safety and Helpfulness**: The RLHF approach also focuses on enhancing the model's safety and helpfulness, making it more reliable for interaction in sensitive domains like healthcare.

- **Clinical Trial Data Analysis**: In the context of the NLI4CT project, Llama-2-7b-chat-hf's capabilities are leveraged to analyze clinical trial reports, a task that requires understanding complex medical terminology and drawing inferences from detailed data.

- **Text Generation for Classification**: The model's fine-tuning in text generation is particularly beneficial for classifying relationships in clinical trial data, such as identifying entailment or contradiction between different sections of the reports.

- **Flexible Prompt Design**: The model allows for customized prompt design, enabling users to guide the model's responses and behavior effectively. This feature is crucial in applications like NLI4CT, where specific and accurate responses are needed.

In summary, Llama-2-7b-chat-hf stands out as a versatile and efficient model for complex NLP tasks, particularly in dialogue and text generation applications. Its balance of scale, training methodology, and computational efficiency makes it a valuable tool in the NLI4CT project and similar endeavors.

2.3 LLAMA Prompt

The Llama 2 prompt format introduces a structured approach to defining the behavior and personality of chat assistants. This format allows for precise control over the system prompt, a crucial aspect in specifying the assistant's role and response style.

```
<s>[INST] <<SYS>>
{{ system_prompt }}
<</SYS>>
```

```
{{ user_msg_1 }} [/INST] {{
    model_answer_1 }} </s><s>[INST
    ] {{ user_msg_2 }} [/INST]
```

For NLI4CT, the system prompt is tailored to guide the model in determining the inference relation between CTRs and statements. This customization is pivotal in achieving accurate and contextually relevant responses.

2.4 Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LORA)

2.4.1 Parameter-Efficient Fine-Tuning

PEFT is a technique designed to fine-tune large language models like LLaMA-2-7B efficiently, especially when computational resources are limited. It focuses on updating a small subset of the model's parameters, reducing the memory and computational overhead typically associated with training large models. This approach is particularly beneficial for users with limited hardware resources, such as a single consumer-grade GPU.

Key aspects of PEFT:

- **Selective Parameter Update**: Instead of updating all parameters, PEFT targets specific layers or parts of the model, such as the attention or feed-forward layers.
- **Resource Efficiency**: By updating fewer parameters, PEFT significantly reduces the memory footprint and computational load, enabling fine-tuning on less powerful hardware.
- **Preservation of Pretrained Knowledge**: PEFT maintains most of the pretrained weights, preserving the general knowledge the model has acquired and focusing on adapting it to specific tasks.

2.4.2 Low-Rank Adaptation

LORA is a technique that introduces low-rank matrices to the model's weights, allowing for efficient adaptation of the model to new tasks. It's particularly effective for fine-tuning large models where traditional fine-tuning methods would be computationally prohibitive.

Key features of LORA:

- **Low-Rank Matrices**: LORA adds low-rank matrices to the weights of certain layers, enabling the model to learn task-specific adaptations without the need for extensive retraining.
- **Computational Efficiency**: The low-rank approach significantly reduces the number of parameters that need to be updated, making it suitable for training on consumer-grade hardware like a single GPU.
- **Flexibility**: LORA can be applied to various parts of the model, such as attention and feed-forward layers, providing flexibility in model adaptation.

2.4.3 Training Setup

Our training setup for LLaMA-2-7B on a single 4090 GPU with torch type `bfloat16` and model loading in 8-bit precision is an excellent example of leveraging advanced techniques to train large models efficiently. Key points of our setup include:

- **Use of bfloat16**: The use of bfloat16 (Brain Floating Point) reduces the memory footprint while maintaining a balance between precision and range. This choice is crucial for handling large models on consumer-grade GPUs.
- **8-bit Model Loading**: Loading the model in 8-bit precision further reduces the memory requirements, enabling you to train a large model like LLaMA-2-7B on a single GPU.
- **Memory Usage**: Our setup's memory consumption of around 20GB with a batch size of 8 is a testament to the efficiency of PEFT and LORA, as well as the precision and loading optimizations.
- **Checkpoint Gradient**: Enabling checkpoint gradient is a smart move to manage memory usage, as it trades off some computational time for reduced memory consumption by recomputing gradients during the backward pass.

In summary, our approach to training LLaMA-2-7B demonstrates an effective use of PEFT and LORA, combined with precision and memory optimizations, to fine-tune a large language model on a single consumer-grade GPU. This setup is a valuable reference for researchers and practitioners working with limited hardware resources.

2.4.4 TRL

We use Transformers Reinforcement Learning (TRL) as our supervised finetuning framework. TRL is a comprehensive library that facilitates the

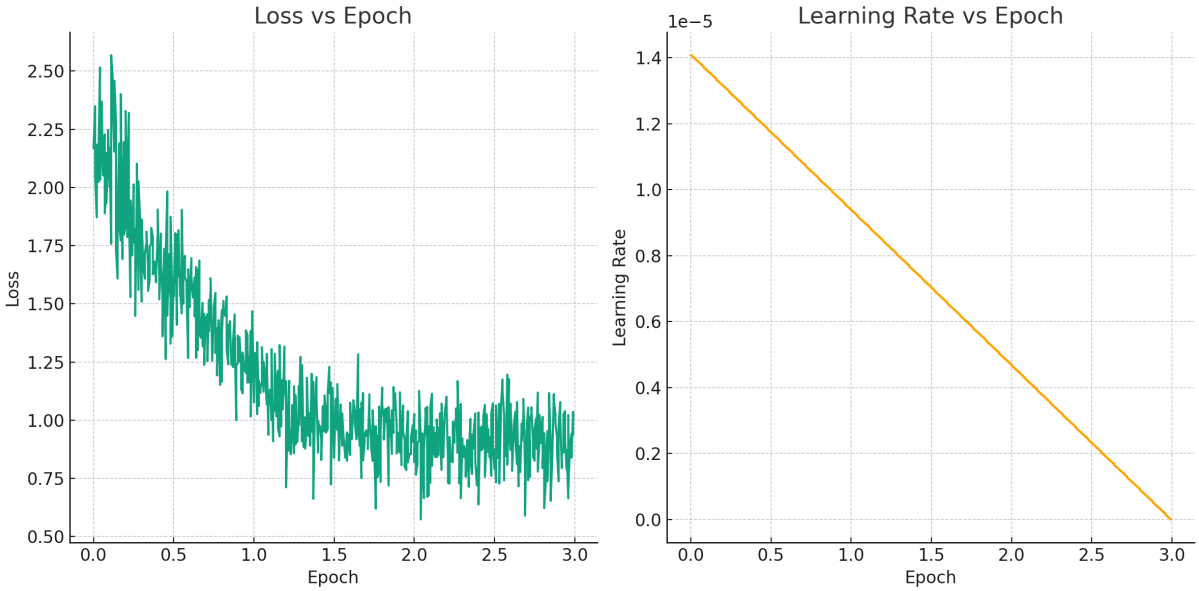

Figure 1: Training progress of the Llama-2-7b-chat-hf model.

training of transformer language models using Reinforcement Learning. TRL integrates seamlessly with the Hugging Face Transformers library and supports various stages of model training, including Supervised Fine-tuning (SFT), Reward Modeling (RM), and Proximal Policy Optimization (PPO). This integration makes TRL a versatile tool for developing advanced NLP models.

2.5 Empirical Results

The input and output structure for the NLI4CT project is designed to mimic a natural conversation flow, where the model receives a prompt (input) and generates a response (output) that classifies the relationship between CTRs and statements as either entailment or contradiction.

2.5.1 Finetune Progress

The figure provides a clear visual representation of how both loss and learning rate evolve over the course of the training epochs.

2.6 Comparative Analysis

2.6.1 Our Results

- **F1 Score (0.667416)**: Indicates a balanced performance between precision and recall, but with potential for improvement.
- **Precision Score (0.675)**: Reflects moderate accuracy in the model's positive predictions.
- **Recall Score (0.660)**: Shows the model's capability in identifying relevant instances, but suggests room for improvement.

2.6.2 Comparison with Other Results

When compared to other submissions in the "SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data," our results show a competitive but not leading performance. For instance:

- **Top Performers (e.g., Zhou et al., 2023; Kanakarajan and Sankarasubbu, 2023)**: Achieved F1 scores around 0.85, significantly higher than our model. These models often employed sophisticated techniques like model ensembling and instruction-finetuned LLMs.
- **Mid-Tier Performers (e.g., Zhao et al., 2023; Feng et al., 2023)**: Had F1 scores in the range of 0.679, closely aligning with our results. These models typically used a combination of advanced NLP models like BioBERT and DeBERTa.
- **Lower-Tier Performers (e.g., Volosincu et al., 2023; Neves, 2023)**: Scored below our model, with F1 scores around 0.64. These models often relied on simpler or less sophisticated approaches.

2.6.3 Conclusion and Future Directions

The "Multi-Evidence Natural Language Inference for Clinical Trial Data" project has demonstrated the potential of LLMs in processing complex medical texts. Our model, while showing a balanced

performance, did not reach the top tier of results, indicating areas for improvement.

The high performance of the leading models underscores the effectiveness of advanced techniques like model ensembling and fine-tuning strategies. Our model's moderate performance suggests that further optimization, possibly through more sophisticated training methods or model architectures, could yield significant improvements.

Future work should focus on exploring these advanced techniques, enhancing the model's precision and recall, and developing robust evaluation methodologies. This will not only improve the model's performance but also ensure its applicability and reliability in critical domains like healthcare. The goal is to bridge the gap between current performance and the top-tier results, ensuring that our model can effectively contribute to the large-scale interpretation and retrieval of medical evidence in clinical trials.

3 PubMedBert, BM25 evidence retrieval approach and TF-IDF Vectorization methods and comparison

3.1 Data Splits and packages used

The experimental data was partitioned into training, development (dev), and test sets. We have equally distributed dataset for model analysis and evaluation. The dev set was utilized for fine-tuning and validation, ensuring the model's robustness and preventing overfitting to the training data. The test set, kept separate until the final evaluation, allowed for an unbiased assessment of the system's generalization performance. The PubMedBERT model was accessed through the Hugging Face Transformers library. The BM25 Evidence Retrieval utilized the BM25Okapi class from the Gensim library. TF-IDF vectorization was performed using the scikit-learn libraries.

3.2 Preprocessing and Hyperparameter Tuning

For the PubMedBERT-based approach, the biomedical statements were tokenized and embedded using the PubMedBERT model. The embeddings were then used to calculate cosine similarity scores. The threshold for entailment prediction was empirically set at 0.12. For BM25 Evidence Retrieval, the primary and secondary trial sections were tokenized, and BM25 scores were calculated for each section against the statement. An arbitrary BM25 score threshold of 1 was employed to retrieve relevant evidence. TF-IDF Entailment Prediction involved vectorizing the statements and trial sections using TF-IDF. Cosine similarity scores were computed to determine entailment.

Table 1: Model Performance

Model Name	F1-score
GPT3.5 Turbo	0.058252
llama-2-dev	0.667416
PubMedBert	0.639405
TF-IDF Entailment	0.502415
BM25 Evidence Retrieval	0.322748

3.3 Results

Quantitative Findings: The PubMedBERT model achieved an F1 score of 0.639405, with precision and recall values of 0.508876 and 0.860000, respectively. TF-IDF Entailment Prediction yielded an F1 score of 0.502415 and a precision score of 0.485981. BM25 Evidence Retrieval reported an F1 score of 0.322748, with precision and recall values of 0.422034 and 0.520000. These results underscore the varying degrees of success across different methodologies.

Quantitative Analysis: A quantitative analysis involved ablations and comparisons to discern the effectiveness of different design decisions. The PubMedBERT model exhibited competitive performance, leveraging embeddings to capture nuanced relationships. TF-IDF Entailment Prediction demonstrated moderate success, while BM25 Evidence Retrieval faced challenges, reflecting the intricacies of biomedical language.

Error Analysis: Error analysis delved into understanding system predictions and potential pitfalls. The PubMedBERT model excelled in capturing subtle entailment relationships but faced challenges in cases heavily influenced by contextual nuances. BM25 Evidence Retrieval struggled with precision, leading to potential false positives. TF-IDF Entailment Prediction demonstrated limitations in distinguishing complex entailment relationships.

3.4 GPT-3.5 Turbo

: The GPT-3.5 Turbo model achieved an F1 score of 0.058252, precision score of 1.000000, and recall score of 0.030000 in the Safe Biomedical Natural

Language Inference for Clinical Trials task showing poor performance among the models discussed above.

4 Conclusion

In summary, the experimental setup, external tools, and quantitative findings provide a comprehensive overview of our approach, revealing both successes and challenges in the task of Safe Biomedical Natural Language Inference for Clinical Trials. The models discussed above successfully identifies the contradiction in the provided clinical trial data, demonstrating its capability to understand and analyze complex medical texts.

A References

1. NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports Maël Jullien1 , Marco Valentino3 , Hannah Frost1, 2, Paul O'Regan2 , Donal Landers2 , André Freitas [link].

2. NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, André Freitas [link].

3. June 2023, A Stage Review of Instruction Tuning [link].

4. ChatGPT-3.5-turbo [link].

5. Huggingface-PubMedBert model [link].

6. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing -Microsoft Research [link].

7. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants Guest Editor (s): Joemon M. Jose,8 Emine Yilmaz,9 João Magalhães,10 Pablo Castells,11 Nicola Ferro,12 Mário J. Silva,13 and Flávio Martins14 [link].

8. An Introduction to Information Retrieval Christopher D. Manning Prabhakar Raghavan Hinrich Schütze [link].

9. Understanding TF-IDF in NLP: A Comprehensive Guide [link].

10. Llama 2: Open Foundation and Fine-Tuned Chat Models [link].

11. Paper Reading: Llama 2 — A skip forward for open source NLP [link].