Lisbon Computational Linguists at SemEval-2024 Task 2: Using A Mistral-7B Model and Data Augmentation

Artur Guimarães
INESC-ID and IST
University of Lisbon
Lisbon, Portugal
artur.guimas@gmail.com

Bruno Martins
INESC-ID and IST
University of Lisbon
Lisbon, Portugal
bruno.g.martins@tecnico.ulisboa.pt

João Magalhães
NOVA-LINCS
NOVA University of Lisbon
Lisbon, Portugal
jmag@fct.unl.pt

Abstract

This paper describes our approach to the SemEval-2024 safe biomedical Natural Language Inference for Clinical Trials (NLI4CT) task, which concerns classifying statements about Clinical Trial Reports (CTRs). We explored the capabilities of Mistral-7B, a generalistic open-source Large Language Model (LLM). We developed a prompt for the NLI4CT task, and fine-tuning a quantized version of the model using a slightly augmented version of the training dataset. The experimental results show that this approach can produce notable results in terms of the macro F1-score, while having limitations in terms of faithfulness and consistency. All the developed code is publicly available on a GitHub repository[1].

1 Introduction

Large Language Models (LLMs) currently achieve state-of-the-art performance on different Natural Language Processing (NLP) tasks, including in the assessment of textual entailment relations. However, these models are heavily susceptible to shortcut learning, factual inconsistency, and performance degradation when exposed to data from specialized domains (e.g., medical data).

Noting the aforementioned challenges, Task 2 at SemEval-2024 addressed a safe biomedical Natural Language Inference for Clinical Trials (NLI4CT) task (Jullien et al., 2024), which concerns classifying statements about Clinical Trial Reports (CTRs). NLI4CT investigated the accuracy, faithfulness, and consistency of the reasoning performed by LLMs in this particular medical task, which concerns determining the inference relation (i.e., entailment or contradiction) between CTRs and statements making some type of claim about a single CTR or a pair of CTRs. Given the specific focus on assessing model faithfulness and consistency (i.e., the ability to make correct predictions for the correct reasons), the dataset associated to the task involved the systematic application of controlled interventions, either preserving or inversing the entailment relations originally generated by clinical domain experts. This way, the task investigated the robustness of NLI models in their representation of the semantic phenomena necessary for complex inference in clinical inference settings.

Our approach to the NLI4CT task involved the use of open-source LLMs, with good results in general purpose benchmarks[2] and capable of following task instructions. We opted for `Mistral-7B-Instruct-v0.2`[3] (Jiang et al., 2023), quantizing the model to 4-bits and simultaneously using Low-Rank Adaptation (LoRA) (Hu et al., 2021; Dettmers et al., 2023) to fine-tune the model to the NLI4CT task, using a slightly augmented version of the training dataset that features a mixture of manually curated and synthetic statements.

Our overall best submission to the task achieved a **macro F1-score** of 0.80 (1st place on the leaderboard), a **consistency** score of 0.72 (15th), and a **faithfulness** score of 0.83 (11th). Our method excels in classification accuracy, but fails at being robust to perturbations on the statements, i.e. predicting the same label on contradictory examples and different labels on paraphrased examples.

2 Background

The NLI4CT task concerns inferring if a statement is entailed or contradicted by a given textual context, each statement referring to one or two CTRs. These CTRs belong to a corpus consisting of 1000 different trials concerning breast cancer, extracted from the United States National Library

[1] https://github.com/araag2/SemEval2024-Task2

[2] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

of Medicine[4]. These trial reports are exclusively written in the English language and average 817 words in length.

CTRs are divided into four sections: **Eligibility Criteria**, describing a set of conditions to allow or exclude patients in the trial; **Interventions**, detailing all information about the conducted treatments; **Results**, outlining outcomes and experimental results gathered through the trial; and **Adverse Events**, reporting patient observations concerning symptoms and physiological signs. An instance of the NLI4CT dataset contains either one or two CTRs (i.e., cases denoted as single or comparison, respectively), a statement, a section marker, and a ground-truth label (i.e., entailment or contradiction). An example is shown next.

`Primary Trial:`
INTERVENTION 1:
• Letrozole, Breast Enhancement, Safety.
• Single arm of healthy postmenopausal women to have two breast MRI (baseline and post-treatment). Letrozole of 12.5 mg/day is given for three successive days just prior to the second MRI.

`Secondary Trial:`
INTERVENTION 1:
• FFDM Mammography Exam - LIP Algorithm
• Screening or diagnostic Full Field Digital Mammography (FFDM) exam
INTERVENTION 2:
• FFDM Mammography Exam - SIP Algorithm.
• The same 130 raw data images were externally reprocessed with the Siemens processing algorithm.

`Section:` Intervention

`Statement:` The primary trial and the secondary trial both used MRI for their interventions.

`Label:` Entailment.

The dataset[5] provided by the task organizers considered training, development, practice-test, and test splits (the last two without ground-truth labels during the competition), with a general statistical characterization provided in Table 1.

[4] https://clinicaltrials.gov/

[5] https://github.com/ai-systems/Task-2-SemEval-2024/blob/main/README.md

Set	# Samples	Single - Comparison	Entailment - Contradiction
Training	1700	60.9% - 39.1%	50% - 50%
Development	200	70% - 30%	50% - 50%
Pratice-test	2142	71.2% - 28.8%	34.1% - 65.9%
Test	5500	46.4% - 53.6%	33.5% - 66.5%

Table 1: The NLI4CT task dataset.

The first two splits, i.e. training and development, are similar to those used in the SemEval-2023 edition of the task (Jullien et al., 2023b), based on the work by Jullien et al. (2023a). These are two balanced sets, with mostly unique CTR-statement associations (i.e., statements that are not rephrasing or contradicting other ones). On the other hand, this composition contrasts with the practice-test and test splits, that are both imbalanced and almost solely composed of statements featuring interventions (e.g., paraphrasing, contradicting, or appending text) over a small set of original statements (< 10%), as show in Table 2. This distribution favours systems that focus on robustly classifying a small set of samples.

Set	# Interventions	Preserving - Altering (Label)
Pratice-test	1942 (90.7%)	82.7% - 27.3%
Test	5000 (90.9%)	82.7% - 27.3%

Table 2: Interventions over statements on the test splits.

3 System Overview

We now describe our general approach to the NLI4CT task.

3.1 Choice of LLM

When deciding on how to build our NLI4CT system, we started by testing the zero-shot and few-shot capabilities of several open-source LLMs, before settling on the use of `Mistral-7B-Instruct-v0.2`. In addition to good preliminary results, this model also allowed us to process arbitrarily long input texts, which in this task is relevant since some CTRs can exceed the length of 3000 tokens.

3.2 Model Prompting

A great deal of attention is currently given to prompting techniques, as the successful use of an LLM can be severely impaired by suboptimal prompts, and also since instruction fine-tuning (Chowdhery et al., 2022; Chung et al., 2022) is dependant on prompt quality. In order to address the task of choosing a good prompt, we started by

creating a prompt template that we deemed as suitable for the task at hand, sub-diving our prompt into distinct parts (pre-pended with "$") that can latter be replaced with different textual realizations. The overall structure is illustrated next.

```
$task_description
$ctr_description

Primary Trial:
$primary_evidence

Secondary Trial:
$secondary_evidence

$statement_description
$statement

$option_description
```

Four of the parts are **sample independent**: `$task_description` provides a general description for the natural language inference task between CTRs and statements; `$ctr_description` delineates the general contents of a CTR and its different sections; `$statement_description` conveys the nature of the $statement; and lastly `$option_description` outlines the answers we expect from the model (e.g., an answer of YES or NO, depending on whether the CTR supports the statement). Conversely, `$primary_evidence`, `$secondary_evidence`, and `$statement` are **sample dependent**, as these parts should be replaced by the primary CTR, the secondary CTR (if applicable), and the statement, respectively.

We created 5 base prompts (see Appendix A.1) for each of these 4 sample independent parts, yielding 625 possible combinations for the general template. We evaluated all the combinations on the development set, and chose the prompt that yielded the top **macro F1-score**, which was the following.

<s>[INST]The objective is to examine semantic entailment relationships between individual sections of Clinical Trial Reports (CTRs) and statements articulated by clinical domain experts. CTRs elaborate on the procedures and findings of clinical trials, scrutinizing the effectiveness and safety of novel treatments. Each trial involves cohorts or arms exposed to distinct treatments or exhibiting diverse baseline characteristics.
Comprehensive CTRs comprise four sections: (1) ELIGIBILITY CRITERIA delineating conditions for patient inclusion, (2) INTERVENTION particulars specifying type, dosage, frequency, and duration of treatments, (3) RESULTS summary encompassing participant statistics, outcome measures, units, and conclusions, and (4) ADVERSE EVENTS cataloging signs and symptoms observed. Statements posit claims regarding the information within these sections, either for a single CTR or in comparative analysis of two. To establish entailment, the statement's assertion should harmonize with clinical trial data, find substantiation in the CTR, and avoid contradiction with the provided descriptions.
The following descriptions correspond to the information in one of the Clinical Trial Report (CTR) sections.

Primary Trial:
$primary_evidence

Secondary Trial:
$secondary_evidence

Reflect upon the ensuing statement crafted by an expert in clinical trials.
$statement
Respond with either YES or NO to indicate whether it is possible to determine the statement's validity based on the Clinical Trial Report (CTR) information, with the statement being supported by the CTR data and not contradicting the provided descriptions.[/INST] Answer:

3.3 Generating Answers

With the aforementioned template, we used the Python HuggingFace Transformers library[6] to generate answers with `Mistral-7B-Instruct-v0.2`, using as parameters `do_sample=True`, `top_k=5`, and `max_new_tokens=30`. We opted not to constrain the generation process, instead looking for sets of words, associated to each label, in the sequence of generated tokens. The words "Yes", "yes", and "entailment" were used for the entailment class, while the words "No", "no" and "contradiction" were used for the contradiction class. Preference was given to the first token in the sequence that belongs to either of the sets, and if none were found we label the instance as entailment.

3.4 Data Augmentation

The NLI4CT dataset features 1700 training instances and 200 development instances, which is perhaps insufficient for fine-tuning an LLM in order to generalize to a testing split that is almost

[6] https://huggingface.co/docs/transformers/en/index

thrice as large. We decided to augment the available, and we created the 3 different training splits outlined in Table 3.

Set	# Samples	Single - Comparison	Entailment - Contradiction
Train_Manual	2344	61.8% - 38.2%	50% - 50%
Train_Manual-Synthetic	3720	63.7% - 36.3%	50% - 50%
Train_Full-Synthetic	11011	60.9% - 39.1%	46.3% - 53.7%

Table 3: Results from task data augmentation.

The three new sets were constructed as follows:

- **Train_Manual:** Starting from the train split, we added queries created by using pre-existing samples with the entailment class, negating them using the Python negate library[7] (i.e., to generate corresponding contradiction examples), and also manually paraphrasing the original instance (i.e., to generate different entailment samples). All 644 additional samples that were generated through this procedure were manually curated;

- **Train_Manual-Synthetic:** starting from the **Train_Manual** dataset, we added 1376 new automatically generated instances to this set-Half of the new instances were generated with the negate library, and the other half we generated by paraphrasing existing statements using the `Mistral-7B-Instruct-v0.2` model;

- **Train_Full-Synthetic:** Starting from the train split, we added 9311 new samples, using the negate library on entailment instances, and the `Mistral-7B-Instruct-v0.2` model to paraphrase each original statement 5 times.

3.5 Instruction Fine-tuning

Noting that `Mistral-7B-Instruct-v0.2` is a generalist instruction fine-tuned model, we sought to fine-tune this LLM to the NLI4CT task, using the aforementioned instruction prompt. Due to hardware limitations and the need for handling very long sequences (i.e., up to 6000 tokens), we quantized the model to 4-bit representations of the parameters, and used LoRA to efficiently train the model (Hu et al., 2021). Model training used a supervised fine-tuning objective based on auto-regressive language modelling, completing the input prompt with the correct label for each instance (i.e., outputting either "Yes" or "No" after "Answer:" in the prompt). The implementation relied on the PEFT[8] and TRL[9] Python libraries.

4 Experimental Setup

Making official submissions to the task leaderboard required the participants to submit full runs of the test set, outputting a label for each of its instances. We obtained the labels for each instance by following the procedure described in Subsection 3.3.

The task uses the following evaluation measures: **macro F1-score**, i.e. the arithmetic mean of precision and recall, averaged over the two classes; **Faithfulness**, i.e. a measure created to assess the capacity of model to arrive at the correct prediction for the correct reason, calculated by measuring the ability of model to change its prediction label after semantically altering a statement; and **Consistency**, which completes faithfulness by measuring the ability of a model in outputting the same prediction for semantically equivalent statements[10]. We evaluated our runs using the official metrics obtained from the leaderboard.

Following the training procedure described in Section 3.5, we tested different combinations of training data (as described in Section 3.4). The full set of hyper-parameters associated to our best run can be found in Appendix A.2. All the different runs used Python libraries and packages that can be found in our GitHub repository[11].

5 Results and Discussion

In total, we submitted 38 runs to the leaderboard (e.g., testing different hyper-parameter choices). Table 4 presents our most important results, showing the best result that we achieved with each training set. In turn, Figure 1 compares our overall best run with the top three submissions, per metric.

Trained on which Sets	F1-Score	Faithfulness	Consistency
None (Zero-Shot)	0.67 (3)	0.61 (8)	0.53 (8)
Train	0.81 (2)	0.72 (3)	0.69 (2)
Train_Manual	**0.82 (9)**	0.76 (9)	0.71 (9)
Train_Manual-Synthetic	0.80 (1)	**0.83 (1)**	**0.72 (2)**
Train_Full-Synthetic	0.78 (1)	0.78 (0)	0.71 (0)

Table 4: Results on different training datasets.

[7] https://github.com/dmlls/negate

[8] https://huggingface.co/docs/peft/en/index

[9] https://huggingface.co/docs/trl/en/index

[10] https://github.com/ai-systems/Task-2-SemEval-2024/blob/main/evaluate.py

[11] https://github.com/araag2/SemEval2024-Task2/blob/main/environment.yml

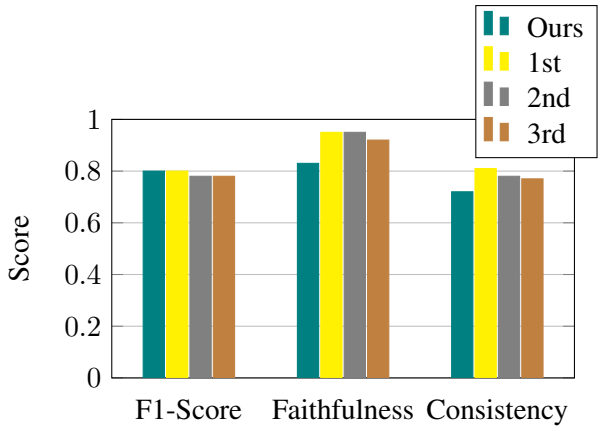

Figure 1: Comparison of top submissions against our system, according to different evaluation metrics.

We tested the LLM without any training (i.e., zero-shot results) and after selecting the best prompt. As expected, there is a significant difference in performance towards fine-tuned models. Overall, a mixture between manually curated samples and synthetically generated ones performed best, outperforming the best run that did not use any data augmentation. If more instances could have been manually curated, specifically targeting adversarial re-writes of the same statements, we theorize results could be improved much further. Even though **Train_Full-Synthetic** corresponds to the largest training set (i.e., featuring 11011 samples), the lack of quality in the automatically generated statements potentially impaired the **F1-score** while also limiting **consistency** and **faithfulness**.

The run trained with **Train_Manual-Synthetic** corresponds to our best overall result. When compared to the top submissions, we can see that our F1-score corresponds to a tie with another system in the 1st place of the leaderboard. However, results are much worse in the other two metrics, with significant differences between the top systems (i.e., with scores of 0.95 in faithfulness and 0.81 in consistency) and our submission,.

Since the ground-truth labels have been released for the test set after the competition, specifying which type of intervention was made in each instance, we are able to analyse our system's errors (see Table 5), to support a discussion on the main short-comings of our system.

Type of Error	# Occurrences / # Total Samples of that Type
Base Statement Errors	99 / 500 (19.8%)
Intervention Errors	1328 / 5000 (26.7%)
Total Errors	1427 / 5500 (25.9%)
Label Preserving Intervention Errors	1177 / 1328 (88.6%)
Label Altering Intervention Errors	151 / 1328 (11.4%)
Paraphrasing Errors	344 / 1500 (22.9%)
Text Appending Errors	609 / 1500 (40.6%)
Contradicting Errors	293 / 1500 (19.5%)
Numerical Paraphrasing Errors	58 / 224 (25.9%)
Numerical Contradicting Errors	24 / 276 (8.7%)

Table 5: Error analysis for our best overall run, categorizing errors by intervention types.

Comparing all errors across the different instance types, the average error rate is much higher on intervention errors (26.7%) against base statement errors (19.8%), which is to be expected as our training sets had less of these examples. Specifically, we can see that label preserving interventions (88.6%) have a high percentage of errors. Our system can identify instances which suffered contradictory interventions with an error rate of 19.5% for textual changes, and 8.7% for numerical changes. Instances that were perturbed with paraphrasing cause an error rate of 22.9%, while numerical paraphrasing errors correspond to 25.9%. At the worst end we have the samples with text appended to the end, which causes an error rate of 40.6%. Note that we did did not explicitly augment the training instances by appending text to the existing statements, and the absence of examples like this was very costly in terms of the final results.

6 Conclusions

Adapting evaluation methodologies to better inform the safe deployment of LLMs in critical domains is an urgent necessity. The NLI4CT task at SemEval-2024 addressed this specific concern, and through our participation we improved our understanding on how LLMs can be fine-tuned to encompass robust results on clinical natural language inference. For future work we would like to explore the following ideas:

- Test our general approach with different models, specifically considering models fine-tuned in the medical domain (e.g., `qCammel-70-x`[12] or BioMistral[13]);
- Refining the considered prompt through prompt recently-proposed optimization methods (Wen et al., 2023; Guo et al., 2023), instead of relying on manually curated prompts;

[12] https://huggingface.co/augtoma/qCammel-70-x
[13] `https://huggingface.co/BioMistral`

- Incorporating additional training data, e.g. by generating a more diverse set of instances from the CTR data made available in the context of other shared tasks (e.g., the CTR data from the Text Retrieval Conference (TREC) clinical trials track[14]);

- Carefully curating a new training set, with a focus on statement interventions rather than quantity of base statements, in order to better guide the model into understanding the nuances of textual and numerical paraphrasing/contradiction.

Acknowledgments

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by Fundação para a Ciência e Tecnologia (FCT), through the project with reference UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020).

References

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized llms. *arXiv:2305.14314*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv:2309.08532*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv:2310.06825*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

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

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv:2302.03668*.

A Appendix

We now present additional details about the prompt considered for instructing the `Mistral-7B-Instruct-v0.2` model, and about the hyper-parameters considered for model fine-tuning.

[14] https://www.trec-cds.org/

A.1 Base Descriptions For Each Prompt Part

This section presents the five different alternatives that were considered for the different parts of the `Mistral-7B-Instruct-v0.2` prompt.

A.1.1 Task Description Part

1 : Consider the task of determining semantic entailment relations between individual sections of Clinical Trial Reports (CTRs) and statements made by clinical domain experts. Note that CTRs outline the methodology and findings of a clinical trial, which are conducted to assess the effectiveness and safety of new treatments. Each trial involves 1-2 patient groups, called cohorts or arms, and these groups may receive different treatments, or have different baseline characteristics. The complete CTRs contain 4 sections, corresponding to (1) a list of the ELIGIBILITY CRITERIA corresponding to the conditions for patients to be allowed to take part in the clinical trial, (2) a description for the INTERVENTION that specifies the type, dosage, frequency, and duration of treatments being studied, (3) a summary of the RESULTS, detailing aspects such as the number of participants in the trial, the outcome measures, the units, and the conclusions, and (4) a list of ADVERSE EVENTS corresponding to signs and symptoms observed in patients during the clinical trial. In turn, the statements are sentences that make some type of claim about the information contained in one of the aforementioned sections, either considering a single CTR or comparing two CTRs. In order for the entailment relationship to be established, the claim in the statement should be related to the clinical trial information, it should be supported by the CTR, and it must not contradict the provided descriptions.

2 : You are tasked with determining support relationships between individual sections of Clinical Trial Reports (CTRs) and clinical statements. CTRs detail the methodology and findings of clinical trials, assessing effectiveness and safety of new treatments. CTRs consist of 4 sections: (1) ELIGIBILITY CRITERIA listing conditions for patient participation, (2) INTERVENTION description specifying type, dosage, frequency, and duration of treatments, (3) RESULTS summary detailing participants, outcome measures, units, and conclusions, and (4) ADVERSE EVENTS listing signs and symptoms observed. Statements make claims about information in these sections, either for a single CTR or comparing two.

3 : Evaluate the semantic entailment between individual sections of Clinical Trial Reports (CTRs) and statements issued by clinical domain experts. CTRs expound on the methodology and outcomes of clinical trials, appraising the efficacy and safety of new treatments. The statements, on the other hand, assert claims about the information within specific sections of CTRs, for a single CTR or comparative analysis of two. For entailment validation, the statement's claim should align with clinical trial information, find support in the CTR, and refrain from contradicting provided descriptions.

4 : The objective is to examine semantic entailment relationships between individual sections of Clinical Trial Reports (CTRs) and statements articulated by clinical domain experts. CTRs elaborate on the procedures and findings of clinical trials, scrutinizing the effectiveness and safety of novel treatments. Each trial involves cohorts or arms exposed to distinct treatments or exhibiting diverse baseline characteristics. Comprehensive CTRs comprise four sections: (1) ELIGIBILITY CRITERIA delineating conditions for patient inclusion, (2) INTERVENTION particulars specifying type, dosage, frequency, and duration of treatments, (3) RESULTS summary encompassing participant statistics, outcome measures, units, and conclusions, and (4) ADVERSE EVENTS cataloging signs and symptoms observed. Statements posit claims regarding the information within these sections, either for a single CTR or in comparative analysis of two. To establish entailment, the statement's assertion should harmonize with clinical trial data, find substantiation in the CTR, and avoid contradiction with the provided descriptions.

5 : Consider the problem of assessing semantic entailment connections between distinct sections of Clinical Trial Reports (CTRs) and statements put forth by clinical domain experts. To establish entailment, the statement's assertion should be supported from the CTR, not contradicting the provided descriptions. In brief, CTRs elucidate the procedures and findings of clinical trials, evaluating the efficacy and safety of emerging treatments. Complete CTRs encompass four sections: (1) ELIGIBILITY CRITERIA specifying conditions for patient inclusion, (2) INTERVENTION details on the type, dosage, frequency, and duration of treatments, (3) RESULTS summarizing the participant statistics, outcome measures, units, and conclusions, and (4) ADVERSE EVENTS listing observed signs and symptoms. Statements advance claims about the information within these sections, either for a single CTR or in a comparative analysis of two CTRs.

A.1.2 CTR Description Part

1 : The following descriptions correspond to the information in one of the Clinical Trial Report (CTR) sections.

2 : The provided descriptions coincide with the content in a specific section of Clinical Trial Reports (CTRs), detailing relevant information to the trial.

3 : The provided descriptions correspond to the content found in one of the four standard clinical trial report sections.

4 : The provided descriptions pertain to the contents found within one of the sections of Clinical Trial Reports (CTRs).

5 : The descriptions that follow correspond to the information contained in one of the standard sections of the clinical trial reports.

A.1.3 Statement Description Part

1 : Consider also the following statement generated by a clinical domain expert, a clinical trial organizer, or a medical researcher.

2 : Contemplate the ensuing statement formulated by a clinical expert or researcher.

3 : Review the subsequent statement provided by an expert in clinical trials, attending to the medical terminology and carefully addressing any ambiguities.

4 : Deliberate upon the subsequent statement formulated by an healthcare practitioner, a coordinator of clinical trials, or a medical researcher.

5 : Reflect upon the ensuing statement crafted by an expert in clinical trials.

A.1.4 Option Description Part

1 : Answer YES or NO to the question of whether one can conclude the validity of the statement with basis on the clinical trial report information.

2 : Indicate with either YES or NO whether it is possible to determine the validity of the statement based on the Clinical Trial Report (CTR) descriptions. An answer of YES means that the statement is supported by the CTR descriptions, not contradicting the provided information.

3 : Provide a YES or NO response indicating if it's possible to assess the statement's validity based on the information presented in the clinical trial report descriptions. Do this by interpreting the medical terminology and the context in both the report and the statement, carefully addressing any ambiguities or gaps in the provided information.

4 : Respond with either YES or NO to indicate whether it is possible to determine the statement's validity based on the Clinical Trial Report (CTR) information, with the statement being supported by the CTR data and not contradicting the provided descriptions.

5 : Indicate with a YES or NO response whether it is possible to assess the statement's validity based on the clinical trial report data.

A.2 Full List of Hyper-Parameters

The full list of hyper-parameters considered for model fine-tuning can be seen in the source-code in our GitHub repository[15].

The chosen parameters concerning model quantization options are as follows.

```
load_in_4bit = True
bnb_4bit_quant_type = "nf4"
bnb_4bit_compute_dtype = torch.bfloat16
```

[15] https://github.com/araag2/SemEval2024-Task2/blob/main/finetune_Mistral.py

```
bnb_4bit_use_double_quant = False
```

The parameters concerning the use of Low-Rank Adaptation (LoRA) are as follows.

```
lora_r = 64
lora_dropout = 0.1
lora_alpha = 16
bias = "none"
```

Finally, the general model training hyper-parameters are as follows.

```
train_epochs = 5
batch_size = 2
gradient_accumulation_steps = 4
learning_rate = 2e-5
pooling = "mean"
max_sequence_length = 6000
```