BpHigh at SemEval-2023 Task 7: Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data?

Bhavish Pahwa
Mindtickle, Pune, India
bhavishpahwa@gmail.com

Bhavika Pahwa
Government medical College, Aurangabad
bhavikapahwa@gmail.com

Abstract

Many nations and organizations have begun collecting and storing clinical trial records for storage and analytical purposes so that medical and clinical practitioners can refer to them on a centralized database over the internet and stay updated with the current clinical information. Clinical trial records have increased significantly, making it difficult for many medical and clinical practitioners to stay updated with the latest information. To help and support medical and clinical practitioners, there is a need to build intelligent systems that can update them with the latest information in a byte-sized condensed format and, at the same time, leverage their understanding capabilities to help them make decisions. This paper describes our contribution to SemEval 2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT). Our results show that there is still a need to build domain-specific models as smaller transformer-based models can be finetuned on that data and outperform foundational large language models like GPT-3.5. We also demonstrate how the performance of GPT-3.5 can be increased using few-shot prompting by leveraging the semantic similarity of the text samples and the few-shot train snippets. We will also release our code and models on open-source hosting platforms, GitHub and HuggingFace.

1 Introduction

The Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) shared task (Jullien et al., 2023) aims at building clinical support systems on top of clinical trial reports to help clinical professionals. Clinical Trials are studies where the researchers test new ways to prevent, diagnose or treat a disease/disorder or find newer ways to improve the quality of life for people with chronic diseases. These studies are available as Clinical Trial Reports (CTRs), providing detailed study methods and results. Clinical decision support systems can aid clinical and medical professionals in their day-to-day tasks and provide meaningful insights to complement their productivity. Clinical decision support systems consist of common frameworks and analytical software built on EHRs(Electronic health reports). However, the growth and development in AI and NLP opens new doors and opportunities which entrepreneurs, researchers, and engineers can leverage to build next-gen intelligent systems which can leverage the contextual information of EHRs and CTRs. Nevertheless, one must be vigilant and take extra precautions while building intelligent systems for clinical domains, as it is a complex domain (Sutton et al., 2020).

Recently many researchers have started building domain-specific versions of famous transformer architectures like BERT, GPT, and T5. Alsentzer et al. (2019) demonstrate how domain-specific Clinical BERT, Clinical BioBERT outperform standard BERT model on MedNLI natural language inference task (Romanov and Shivade, 2018). However, there are certain limitations that such models face, which are primarily due to the data such domain-specific models are trained on. As the data to train clinical models is usually small and usually belongs to a single institution, group of institutions, or a single region, it cannot generalize to the differences in clinical practice in other institutions and regions. There is a need for creating more datasets in the clinical NLP domain and also for consideration to be given to including data from different regions and institutions. Creating large multi-lingual datasets containing clinical data of multiple geographies can help create more robust and generalized models.

Many researchers also believe that more stress should be given to building generalized large language models trained on a large corpus of open data that performs well in the clinical domain instead of pre-training and fine-tuning domain-specific models. After the release of GPT-3 (Brown et al., 2020),

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1936–1944

many researchers have also tried to experiment and observe whether GPT-3 can be used for clinical systems and how it might benefit clinical practitioners and patients (Korngiebel and Mooney, 2021). The most significant pitfalls they find are the interpretability and hallucination in generative models.

Many countries have also realized why it is necessary to have clinical datasets and store clinical trial records digitally that can represent the experiments on diseases and illnesses more relevant to their population.[1] The NLI4CT shared task also uses one such database by the U.S. National Library of Medicine with clinical trial records of clinical investigations carried out globally by private and public entities.[2] The organizers of the task have created a set of breast cancer clinical trial records, statements, explanations, and labels that have been annotated by experts who belong to the clinical field. The NLI4CT shared task consists of two subtasks, subtask 1 is a textual entailment task, whereas subtask 2 is a retrieval-based task.

We participate in subtask 1 only and demonstrate how cross-encoder models fine-tuned on small training datasets can outperform GPT-3.5 (Ouyang et al., 2022) in zero-shot settings. We fine-tune several models based on sentence transformer and cross-encoder architecture for sentence pair modeling. We also compare our models with GPT-3.5 Davinci (text-davinci-003 according to the OpenAI platform) in zero-shot and few-shot settings. We also show the effect freezing of base transformer layers has while training cross-encoders and on their performance. We will release all our code on GitHub [3] and fine-tuned models on HuggingFace.[4]

2 Dataset Description

The primary dataset consists of a set of CTRs(clinical trial reports) condensed into 4 sections. The subtask-specific dataset for subtask 1 consists of a set of statements which are sentences that act as the premise and are based on the hypothesis, which consists of information present in a particular section of 1 CTR or a comparison of 2 CTRs. The subtask 1 dataset also provides a label whether a given statement-hypothesis pair is an entailment or contradiction.

We explain the 4 sections present in the primary dataset below with the help of the clinical trial report with id- NCT00003782 :-

1. **Eligibility Criteria**:- It acts as a guard which decides which subjects to enroll in the study based on its shoulder angels the inclusion and exclusion criteria. The inclusion criteria – lays down statements, and volunteers who fit in those statements are chosen for the study. The exclusion criteria – provide statements that, if present, eliminate the volunteers from participating in the study.

   For example, in the above-mentioned report, the following inclusion criteria are used- "No ulceration, erythema, infiltration of the skin or underlying chest wall (complete fixation), peau d'orange, or skin edema of any magnitude(Tethering or dimpling of the skin or nipple inversion allowed) " Suppose the volunteer is found to have ulceration /erythema/ infiltration/ edema/ peau d' orange over skin or chest wall infiltration. In that case, the Breast cancer is no longer Stage III A and has already progressed to Stage III B or a higher stage. Dimpling of the skin or nipple inversion is allowed as it is not considered skin involvement and hence will not change the staging.

2. **Intervention**:- This is the active interference done by the researcher, which can be in the form of giving a drug / carrying out a procedure, or implementing preventive measurements.

   For example, in the above-mentioned report, intervention is done by giving chemotherapy using drugs like Doxorubicin, Cyclophosphamide, and Docetaxel. These anticancer drugs act during different stages of cell division to stop cancer cells from proliferating. Doxorubicin and Cyclophosphamide act by destabilizing DNA structure. Docetaxel affects the microtubular system in cells, which helps in equal DNA distribution during cell division.

3. **Result**:- This is the outcome the study intended to achieve through intervention. Due to the different interventions received by the study subjects, different Results were obtained. These are extensively analyzed to determine their Significance. The significant

[1] https://ctri.nic.in/Clinicaltrials/login.php
[2] https://clinicaltrials.gov/ct2/home
[3] https://github.com/bp-high/NLI4CT-Shared-Task-BpHigh
[4] https://huggingface.co/bpHigh

results are used for further practical purposes, and the best methods of Interventions are adapted for medical practice.

For example, in the above-mentioned report, the 8-year survival was maximum (83 %) in subjects who received Doxorubicin with Cyclophosphamide followed by Docetaxel as chemotherapy.

4. **Adverse Events**:- These are usually the unwanted changes that can occur in a subject's health or lab reports during the study timeframe, which may or may not be directly related to the study and study intervention.

   For example, serious adverse effects like Febrile Neutropenia, thromboembolic events, etc., were reported in the above-mentioned report. Febrile Neutropenia is Fever with lab reports suggestive of reduced neutrophil counts, cells protecting the body against infections. Thromboembolic events include blood clot formation in deep veins and complications that may arise due to it.

3 Related Work

Romanov and Shivade (2018) published the MedNLI natural language inference task which kick-started modern research in medical/clinical NLI domain. The dataset consists of the medical history of patients, which has been annotated for NLI tasks by experts in the clinical domain. The authors also present methods to leverage other open-domain NLI datasets and how to incorporate domain knowledge from other external open-domain data. They also developed a baseline model, which achieved an accuracy of 73.5% on the test dataset of the task. The paper also illustrates how baseline model performance can be boosted by using domain-specific word embeddings. In addition, they also introduce a technique to incorporate domain ontologies during the training process of models.

Reimers and Gurevych (2019) released a paper defining Sentence BERT architecture which was built by modifying BERT. The approach involves utilizing Siamese and triplet network structures on top of BERT network to generate sentence embeddings that carry significant semantic information. These embeddings can be compared using cosine similarity. This significantly reduced the run-time while maintaining similar accuracy for sentence-pair regression tasks.

Lewis et al. (2020) conducted a large scale study that covers a wide range of 18 established biomedical and clinical NLP tasks, aiming to identify the effectiveness of various commonly used open-source biomedical and clinical NLP models in diverse settings. They also present their own pre-trained models based on RoBERTa-base and RoBERTa-large where they used a domain-specific vocabulary. Their domain-specific vocabulary is a BPE(Byte-pair encoding) dictionary learned over the pre-training corpus from PubMed. They demonstrated that although their models could not outperform BioBERT (Lee et al., 2019) on biomedical tasks but they were able to beat it on clinical tasks.

4 Methodology

This section discusses our approaches for tackling subtask 1 of the NLI4CT task. For all further references, whenever we refer to a model which has been fine-tuned by using the NLI4CT subtask 1 training dataset, we prepend 'NLI4CT-' to the pretrained model's name. Using the Sentence Transformers package [5], we fine-tune one sentence-transformer model and multiple cross-encoder models. We also compare the performance of our fine-tuned models with GPT-3.5 Davinci (text-davinci-003 according to the OpenAI platform) in zero-shot and few-shot settings.

4.1 Sentence Transformers based approach

We finetune the BioSimCSE-BioLinkBERT-BASE (Kanakarajan et al., 2022)[6] sentence-transformers model on NLI4CT subtask 1 train dataset and name it NLI4CT-BioSimCSE-BioLinkBERT-BASE. We use this pretrained model as it is a SOTA(state of the art) model on semantic textual similarity tasks. We notice that the training data provided to us is relatively small and thus fine-tuning sentence-transformer models might not be the right choice and we could leverage pretrained cross-encoders which often achieve higher performance on sentence-pair classification tasks. (Thakur et al., 2021) Figure 1 shows how sentence-transformers can be used for sentence-pair classification tasks. The hyperparameters for this approach are detailed in Appendix A.

[5] https://github.com/UKPLab/sentence-transformers

[6] https://huggingface.co/kamalkraj/BioSimCSE-BioLinkBERT-BASE

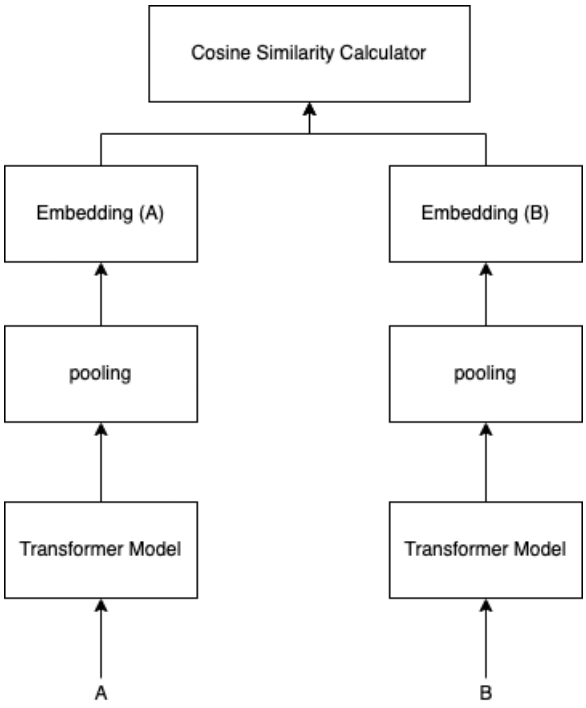

Figure 1: Sentence Transformer

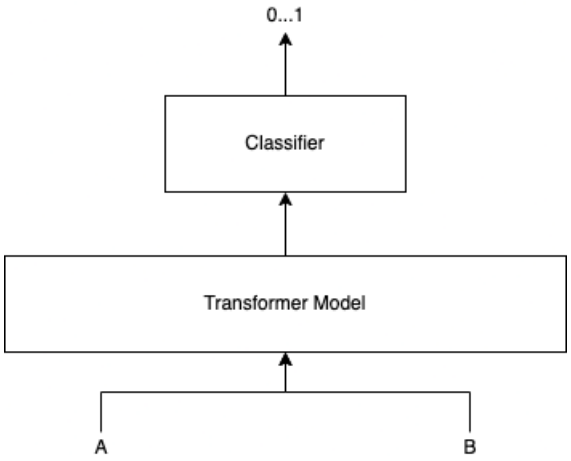

Figure 2: Cross-Encoders

Figure 3: Format of the zero-shot prompt for NLI4CT subtask 1

4.2 Cross-Encoders based approach

We train 6 cross-encoder models on NLI4CT subtask 1 train data using two approaches and based on three pretrained models. We utilise the following pretrained models:-

1. **BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext** (Gu et al., 2021) [7]
2. **BiomedNLP-PubMedBERT-large-uncased-abstract** (Gu et al., 2021) [8]
3. **BioLM-RoBERTa-base-PM-M3-Voc-distill-align** (Lewis et al., 2020) [9]

We train these cross-encoders using two approaches. In the first approach, the models are trained with no layers of the cross-encoder frozen. In the second approach, we freeze all the layers except the cross-encoder pooler and classifier layers. Figure 2 shows how cross-encoders can be leveraged for sentence-pair classification tasks. The hyperparameters for this approach are detailed in Appendix B.

4.3 GPT-3.5 Davinci

We use two approaches while checking the performance of the **GPT-3.5 Davinci model(text-davinci-003)**.

In the first zero-shot approach, we formulate our test dataset's statement and hypothesis as an NLI question and an instruction to be sent to the model as prompt input through OpenAI API. Figure 3 shows how the zero-shot prompt looks.

In the few-shot approach, we first modify all our train dataset samples into the snippet format, as shown in Figure 4. Then we encode all modified train snippets using the BioSimCSE-BioLinkBERT-BASE sentence transformer model and covert them into embeddings to create an embedding dataset. Now for every test sample, we first find the top two train snippets most semantically similar to the statement in the test sample by calculating the cosine similarity between the embedding vectors of the test sample and all candidate train snippets. We then convert the whole snippets plus the test sample into the few-shot prompt format shown in Figure 5. Appendix C contains more details about the instructions, train snippet format, answer format.

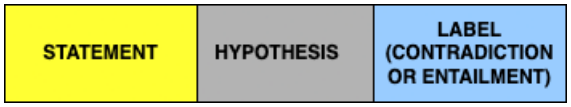

Figure 4: Format of the train snippets to be created for few shot samples

[7] https://huggingface.co/microsoft/
[8] https://huggingface.co/microsoft/
[9] https://github.com/facebookresearch

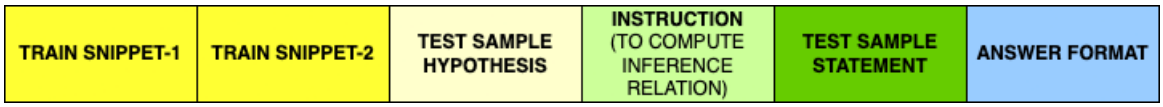

Figure 5: Format of the few-shot prompt

Approach	Macro F1-score Dev	Macro F1-score Test
NLI4CT-BioSimCSE-BioLinkBERT-BASE	0.613	0.595
NLI4CT-BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext(Unfrozen)	0.669	0.655
NLI4CT-BioLM-RoBERTa-base-PM-M3-Voc-distill-align(Unfrozen)	0.694	0.679
NLI4CT-BiomedNLP-PubMedBERT-large-uncased-abstract(Unfrozen)	**0.719**	0.690
NLI4CT-BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext(layers frozen)	0.661	0.646
NLI4CT-BioLM-RoBERTa-base-PM-M3-Voc-distill-align(layers frozen)	0.632	0.615
NLI4CT-BiomedNLP-PubMedBERT-large-uncased-abstract(layers frozen)	0.659	0.637
GPT-3.5 Davinci(text-davinci-003) Zero Shot	0.685	0.664
GPT-3.5 Davinci(text-davinci-003) Few Shot	0.708	**0.694**

Table 1: Results of all the approaches on test and dev dataset

5 Results

The results for all approaches on the Dev and Test dataset are presented in Table 1. The evaluation metric for NLI4CT subtask 1 is the Macro-F1 score. Although the GPT-3.5 Davinci model outperforms our finetuned cross-encoders in a few-shot setting, it is beaten by our finetuned models in a zero-shot setting. We also observe that our approach for choosing semantically similar few-shot train snippets gives a significant performance boost over the zero-shot approach. Another interesting observation comes in the form of the general trend of performance drop when cross-encoder models are trained after freezing the layers. Although freezing the transformer layers and leaving just the classifier and pooler layers unfrozen gives an excellent boost to training time and reduces memory usage, it does not give any performance-based benefits even after training for a good range of iterations and optimizing hyperparameters.

6 Conclusion

We map the textual entailment task as a sentence-pair classification task, and we study the literature to find all approaches that can be used to solve this task. We also study clinical literature to grasp our dataset and understand the definition of whereas properties properly. We also try to find pretrained models already trained on some clinical or biomedical domain that can be finetuned further for our task. We agree on specific approaches/pretrained models and finetune our models based on those pretrained models. To get a good grasp of the performance of our models, we leverage very large language models and compare performance with GPT-3.5 Davinci state-of-the-art models in zero-shot and few-shot settings. We devise a few-shot strategy based on semantic similarity to find the top two few-shot train snippets for each test sample.

We observe that finetuned cross-encoders perform well compared to zero-shot GPT-3.5, and some models also perform comparably to the few-shot GPT-3.5. We also observe that freezing the base transformers' layers while training cross-encoders considerably affects model performance on downstream tasks. Further, we would explore how to increase our model performance and achieve similar performances as unfrozen models by freezing specific layers while training cross-encoders to see their effects. We would also experiment to train cross-encoders based on more pretrained models like BioLinkBERT(Yasunaga et al., 2022).

7 Acknowledgements

We express our appreciation to the Program Committee and the reviewers for their valuable comments and feedback.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Mael 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*.

Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. BioSimCSE: BioMedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *npj Digital Medicine*, 4(1):93.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

A Training details for the Sentence Transformers based approach

Our best performing fine-tuned Sentence Transformer model is trained using the Contrastive Loss function (Hadsell et al., 2006), the evaluator used is the EmbeddingSimilarityEvaluator[10] and the hyperparameters are present in table 2.

[10] https://www.sbert.net

MODEL PARAMETERS	VALUE
Fixed Parameters	
Scheduler	WarmupLinear
Optimizer	AdamW
Weight Decay	0.01
Tuned Parameters	
Num Epochs	5
Warmup Steps	427
Learning Rate	0.0001
Batch Size	2

Table 2: Hyperparameters for Sentence Transformer Approach

B Training details for the Cross-Encoders based approach

We fine-tune all the cross encoder models using the CEBinaryClassificationEvaluator[11] as it maximizes the F1 and use the Binary cross-entropy with logits loss function. The hyperparameters of the respective models are detailed in table 3.

C Details related to zero-shot and few-shot prompt format of GPT 3.5

We present a detailed sample of the zero-shot prompt in Figure 6 along with the respective segments mentioned alongside it.

We also present a sample example train snippet in Figure 7.

The Few-Shot prompt in format is formed by combining the top 2 train snippets, which would be of similar formats as represented in Figure 7, and the test example in the form shown in the zero-shot sample in Figure 6.

[11] https://www.sbert.net/docs

MODEL PARAMETERS	VALUE
Fixed Parameters	
Scheduler	WarmupLinear
Optimizer	AdamW
Weight Decay	0.01
Tuned Parameters	
NLI4CT-BiomedNLP-PubMedBERT-base(Unfrozen)	
Num Epochs	5
Warmup Steps	427
Learning Rate	2e-05
Batch Size	2
NLI4CT-BiomedNLP-PubMedBERT-large(Unfrozen)	
Num Epochs	4
Warmup Steps	342
Learning Rate	1e-04
Batch Size	2
NLI4CT-BioLM-RoBERTa-base(Unfrozen)	
Num Epochs	6
Warmup Steps	512
Learning Rate	2e-04
Batch Size	2
NLI4CT-BiomedNLP-PubMedBERT-base(frozen)	
Num Epochs	5
Warmup Steps	427
Learning Rate	1e-05
Batch Size	2
NLI4CT-BiomedNLP-PubMedBERT-large(frozen)	
Num Epochs	10
Warmup Steps	854
Learning Rate	5e-05
Batch Size	5
NLI4CT-BioLM-RoBERTa-base(frozen)	
Num Epochs	8
Warmup Steps	684
Learning Rate	2e-05
Batch Size	2

Table 3: Hyperparameters for Cross-Encoder Approach

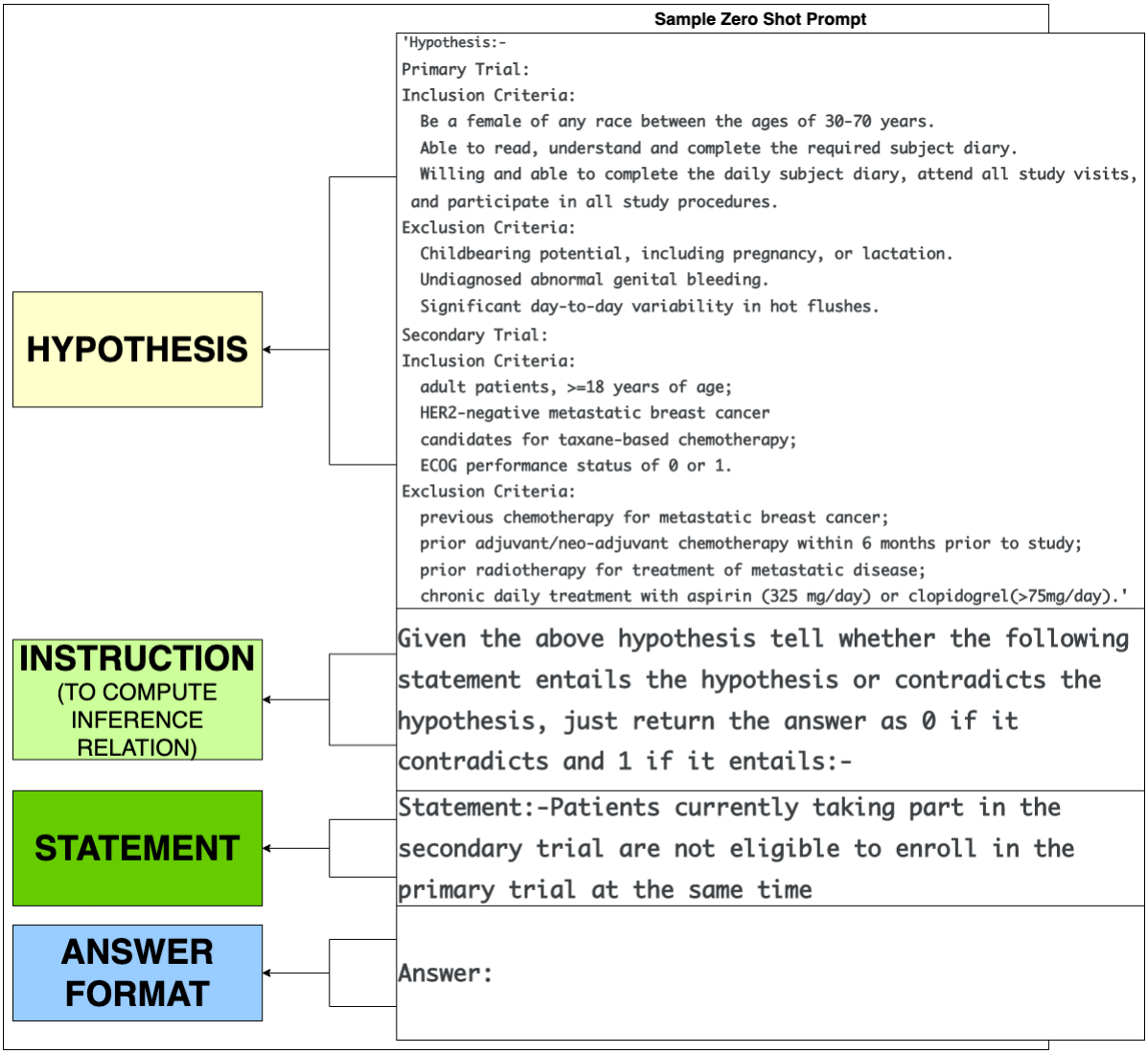

Figure 6: A sample of the zero-shot prompt for NLI4CT subtask 1

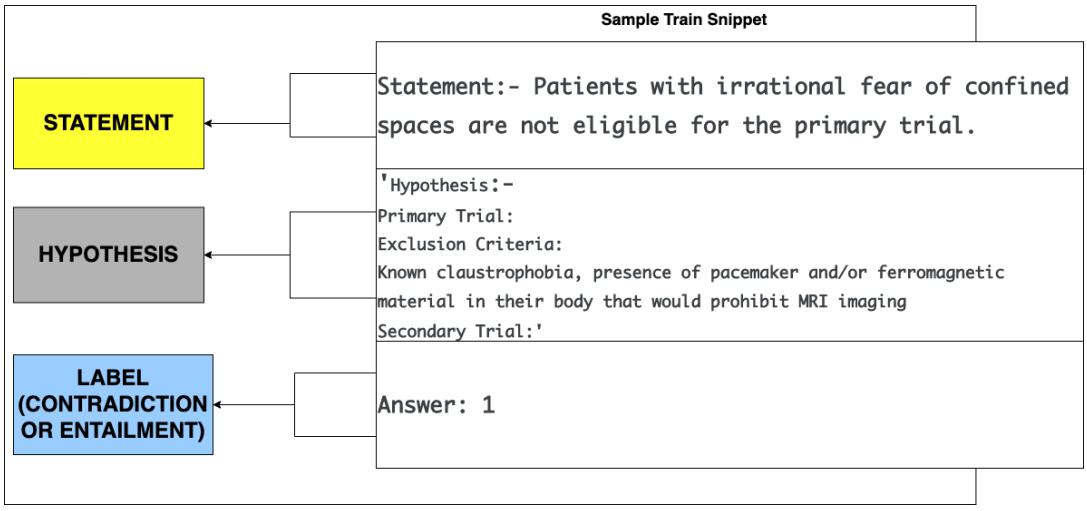

Figure 7: A sample train snippet