BD-NLP at SemEval-2024 Task 2: Investigating Generative and Discriminative Models for Clinical Inference with Knowledge Augmentation

Shantanu Nath
University of Trento, Italy
shantanu.nath@studenti.unitn.it

Ahnaf Mozib Samin
Queen's University, Canada
ahnaf.samin@queensu.ca

Abstract

Healthcare professionals rely on evidence from Clinical Trial Records (CTRs) to devise treatment plans. However, the increasing quantity of CTRs poses challenges in efficiently assimilating the latest evidence to provide personalized evidence-based care. In this paper, we present our solution to the SemEval- 2024 Task 2 titled "Safe Biomedical Natural Language Inference for Clinical Trials". Given a statement and one/two CTRs as inputs, the task is to determine whether or not the statement entails or contradicts the CTRs. We explore both generative and discriminative large language models (LLM) to investigate their performance for clinical inference. Moreover, we contrast the general-purpose LLMs with the ones specifically tailored for the clinical domain to study the potential advantage in mitigating distributional shifts. Furthermore, the benefit of augmenting additional knowledge within the prompt is examined in this work. Our empirical study suggests that DeBERTa-lg, a discriminative task-specific natural language inference model, obtains the highest F1 score of 0.77 and consistency score of 0.76 on the test set, securing the fourth rank on the leaderboard. Intriguingly, the augmentation of knowledge yields subpar results across most cases.

1 Introduction

Clinical trials are conducted on human subjects to test the safety and effectiveness of the medicine prior to designing a new treatment, especially in evidence-based treatments (Avis et al., 2006). Medical professionals prescribe and treat their patients based on clinical trial reports (CTR) in which the methodology and results of clinical trials are outlined. However, the increasing quantity of CTRs poses a challenge for healthcare professionals to manually assess all of them since this process is both time-consuming and labor-intensive (Bastian et al., 2010; DeYoung et al., 2020).

To tackle the aforementioned issue, recent advancements in natural language processing (NLP) encourage medical professionals to employ large language models (LLMs) to interpret and retrieve medical evidence from large quantities of CTRs (Lee et al., 2020). Employing natural language inference (NLI) and textual entailment in the clinical domain (Bowman et al., 2015), professionals can formulate a prompt or statement, for example, *"A minimum bodyweight of 55kg is required to participate in the primary trial."* and input it along with the CTRs into an LLM to determine whether the statement entails or contradicts the evidence from CTRs (Jullien et al., 2023b). Additionally, LLMs can aid in retrieving evidence from the vast amount of CTRs. This technology has the potential to ensure a higher level of precision and efficiency in delivering personalized evidence-based care.

While LLMs have demonstrated significant performance in numerous NLP tasks in recent years (Brown et al., 2020), it is still challenging for them to be deployed in the clinical domain due to their limitation in semantic and quantitative reasoning in language understanding. Moreover, the distribution shift in the clinical domain makes it even more intricate, which requires extensive research in this field (Miller et al., 2020). To address the challenges, Jullien et al. (2024) organizes the SemEval-2024 Task 2, titled "Safe Biomedical Natural Language Inference for Clinical Trials", which aims to investigate the robustness of NLI models when applied to clinical trials with cancer patients. This task seeks to develop an NLI system to connect the new evidence and infer the knowledge from CTRs to find an inferential relation, namely either entailment or contradiction, between a clinical trial document and a statement/claim.

In this study, we investigate the performance of discriminative and generative transformer-based LLMs in the realm of clinical inference. In addition, we explore the potential of clinical domain-specific

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1302–1308

LLMs and compare them with the general-purpose ones with the hypothesis that LLMs pre-trained on clinical data may exhibit superior performance. Furthermore, we study the impact of augmenting knowledge on the semantic reasoning abilities of these LLMs. Our extensive experiments demonstrate that our system, leveraging the discriminative, general-purpose DeBERTa-lg NLI model, achieves an F1 score of 0.77 and consistency score of 0.76 without employing knowledge augmentation on the test set and ranks fourth on the official leaderboard.

2 Background

2.1 Task Definition

In the textual entailment identification task, each input data contains a medical statement, a section name indicating which section the statement claims about, and one or two CTR records that serve as evidence to verify the statement. If the statement only makes claims about one certain trial defined as a primary trial, then only the primary trial will be used as input data. On the other hand, if the statement claims a comparison between a primary trial and a second trial defined as a secondary trial, both CTRs need to be considered as input text. Table 1 presents an example of CTR with four sections. The task is to determine the inferential relation between the medical statement and the associated section in the CTR(s). There are two possible inferential relations for each statement: entailment and contradiction. Models are designed to predict whether each statement entails or contradicts the associated section from the claimed CTR(s).

2.2 Dataset

The Natural Language Inference (NLI) task is designed based on breast cancer CTRs collected by clinical domain experts (Jullien et al., 2023a). Each CTR dataset consists of four sections: intervention, eligibility criteria, results, and adverse events. Each section contains multiple sentences, Formally, $S_t = s_t^1, s_t^2, ..., s_t^n$, here t denotes for the type of section. The participants are provided a text file containing a statement, 1-2 CTRs, an inference label (Entailment or Contradiction), section that is used for the statement. A statement can be made from a single CTR or a comparison between two CTRs.

3 System overview

In this section we describe about our system for this task.

3.1 Input Prompt

According to the data description, a section contains multiple sentences namely evidence. For short input, we consider only the selected sentences of the section which are annotated as related to the statement. All the sentences are concatenated by adding a space in between each sentence to consider a hypothesis. To design the input text, we consider the statement as premise followed by all the selected sentences as claims from the section of the claimed CTR. A separation token, denoted as [SEP], is used between the statement and the claims. For comparison between two claims, we concatenate the selected sentences from both primary and secondary trials, formally, $C^1 s_t^1, .., C^1 s_t^n, ..., C^2 s_t^1, .., C^2 s_t^n$.

3.2 Knowledge Augmentation

In knowledge augmentation, we consider all the sections as evidence. To design input text, we first take all sentences from the related section as a priority. Then, we concatenated the text with other sentences from the rest of the sections. During the comparison between the two trials, we consider the first 500 tokens from the primary trial and 500 tokens from the secondary trial to limit the length of the sentence to 1024 tokens.

In both cases, we design the prompt in the following way:

$$statement\ [SEP]\ primary\ trial:\ C^1 s_t n.$$
$$secondary\ trial:\ C^2 s_n$$

Secondary trials are added only for comparison between two trials.

3.3 Discriminative Models

Discriminative models, in contrast, are focused on learning the decision boundary that separates different classes within the input data. Instead of modeling the entire data distribution, they concentrate on capturing the conditional probability distribution of labels given the input data. We experiment with a collection of transformer-based discriminative pre-trained language models. We choose models that are trained on medical data, such as electronic

Section	Subsection	Sentence
Intervention	N/A	TX/Maintenance Therapy for Stage IIIB/IV Breast Cancer busulfan: Given orally tamoxifen citrate: Given orally
Eligibility	Inclusion	Hepatic function: Bilirubin =< 2 mg% Karnofsky performance status > 60 Creatinine =< 2.0 mg/dl
	Exclusion	Patient is pregnant Are > 100 days from transplant Are on steroids
Results	Outcome Measurement	Event-free Survival Time frame: 11 years
	Results 1	busulfan: Given orally Overall Number of Participants Analyzed: 50
Adverse events	N/A	Total: 2/50 (4.00%) Pulmonary Emboli [2]1/50 (2.00%)

Table 1: An example of a clinical trial record (shortened) containing four sections namely intervention, eligibility, results and adverse events.

health records, biomedical texts, and scientific articles. We used the BioLinkBERT (Yasunaga et al., 2022) model, which was trained on PubMed abstracts along with citation link information. ClinicalBERT (Wang et al., 2023) trained on a large multicenter dataset with a large corpus of 1.2B words of diverse diseases and utilized a large-scale corpus of EHRs from over 3 million patient records to fine-tune the base language model. Bio_ClinicalBERT (Wang et al., 2023), a domain-specific BERT-based model initialized with Bio-BERT model and fine-tuned with electronic health records from ICU patients, namely MIMIC (Johnson et al., 2016). We also choose a task-specific model, proposed by Laurer et al. (2024), based on DeBERTa large and trained on general domain datasets such as MultiNLI (Williams et al., 2018), Fever-NLI (Nie et al., 2019), Adversarial-NLI (ANLI) (Nie et al., 2020), LingNLI (Parrish et al., 2021) and WANLI (Liu et al., 2022) datasets, which comprise 885,242 NLI hypothesis-premise pairs. A classification layer is added on top of the pre-trained layers and fine-tuned on the training set to predict the probability of entailment or contradiction of the statement. An overall descriptions of the models are provided in table 2.

3.4 Generative Models

For comparison, we also solve this task by using generative models. These models use encoder-decoder architecture to encode input text and directly generate output label entailment/contradiction. Similar to discriminative models, we choose SciFive (Phan et al., 2021) as a domain-specific generative pre-trained model that follows The text-to-text transfer transformer (T5) model (Raffel et al., 2019) and sequence-to-sequence encoder-decoder framework. Pubmed and PMC datasets are utilized for training the models and MIMIC is employed to fine-tune for NLI task. To demonstrate the effectiveness of further fine-tuning the Clinical Trial dataset, we apply both zero-shot and few-shot learning approaches on SciFive. On the other hand, we choose Flan-T5 (Chung et al., 2022) as a general domain generative model. Similar to SciFive, Flan-T5 builds based on T5 architecture. For generative models, we slightly change the prompt. For SciFive, $mednli : sentence1 : < premise > sentence2 : < claims >$ and for FlanT5, $natural\ language; inference : premise : < premise > hypothesis : < claims >$.

Type	Model	Parameters	Variation	Task-specific	Domain-specific
Discriminative	ClinicalBERT	110M	base	No	Yes
	BioLinkBERT	340M	large	No	Yes
	BioClinicalBERT	110M	base	No	Yes
	DeBERTa-lg	304M	large	Yes	No
Generative	FlanT5	250M	base	No	No
	SciFive	770M	large	Yes	Yes

Table 2: Model specifics including the number of trainable parameters in million, variation/size, task-specificity (whether further pre-trained on NLI task or not), and domain-specificity (whether pre-trained on medical domain datasets or not) are shown.

Type	Model	W/o knowledge augmentation			With knowledge augmentation		
		Basline F1	Faithfullness	Consistency	Basline F1	Faithfullness	Consistency
Discriminative	ClinicalBERT	0.56	0.35	0.46	0.54	0.35	0.41
	BioLinkBERT	0.57	0.31	0.49	0.57	0.31	0.49
	BioClinicalBERT	0.58	0.47	0.61	0.57	0.31	0.49
	DeBERTa-lg	0.77	0.80	**0.76**	0.75	0.79	0.75
Generative	FlanT5-base	0.58	0.57	0.63	0.51	0.63	0.61
	SciFive (without FFT)	0.49	0.61	0.49	0.47	0.64	0.51
	SciFive (with FFT)	0.44	0.76	0.63	0.65	0.49	0.62

Table 3: Experiment results of the clinical NLI task on the test set for several discriminative and generative models. We report baseline F1, faithfulness, and consistency scores proposed by (Jullien et al., 2024) for each model with/without knowledge augmentation. The best performance with respect to consistency score is bold-faced. Discriminative DeBERTa-lg achieves the best performance while generative models show promise in several cases. Knowledge augmentation implies including all evidence from CTR concatenated with the prompt statement and then passed as input to the LLM. However, Knowledge augmentation shows negligible impact on model performance. FFT stands for further fine-tuning.

4 Experimental setup

We keep the original data split (1700: 200: 5500) provided by the task organizer for training, validation, and testing sets respectively. Huggingface Transformers [1] library is used for tokenization and further finetuning. Data preprocessing steps are mainly adapted from Vladika and Matthes (2023). For short text, the max sequence length for the tokenizer is set to 256 and 512 for long text for BioLinkBERT, ClinicalBERT and BioClinicalBERT. For Deberta-large, SciFive and Flan-T5 the max sequence length for the tokenizer is set to 512 and 1024 for short and long input text respectively. We train all language models for 20 epochs and an AdamW (Loshchilov and Hutter, 2019) optimizer is used for optimization with a default learning rate of 5e-6 for discriminative models and 5e-5 for generative models with weight ratio of 0.06 and weight decay of 0.01. The models are evaluated on the validation set after each epoch by using precision, recall, and F1 scores and saved best model based on least evaluation loss. We measure the performance of the models based on Faithfulness and Consistency proposed by (Jullien et al., 2024). Faithfulness measures the ability to predict the output based on the correct reason. Therefore, if semantic reason change in future, models will be able to change its prediction accordingly. On one hand, consistency measures the ability to make same prediction for the semantically equal statements which ensures the semantic preserving in a model.

5 Results and Discussion

The experimental results are presented in Table 3. All results are calculated on the standard test

[1] https://huggingface.co/docs/transformers

set provided by the shared task organizers. The outcome of the NLI model can be binary: either entailment or contradiction. We use the metrics including baseline F1-score, faithfulness, and consistency, proposed by (Jullien et al., 2024) to calculate the performance of the models.

Among discriminative and generative models, we can observe that generative models including FlanT5 and SciFive outperform the discriminative models e.g. ClinicalBERT, BioLinkBERT, and BioClinicalBERT, in terms of faithfulness and consistency. Given the moderate amount of labeled data for this clinical inference task and textual data as input to the model, which is of low-dimensionality in the latent space compared to high-dimensional vision and speech data, this enables generative models to perform well by learning the joint probability distribution of the input features and the class labels. However, the DeBERTa-lg, which is a discriminative model, achieves the highest F1 scores among all discriminative and generative models. This is likely because the task involves simple binary classification, which can be comparatively easily performed by a discriminative model by separating the data points in the data manifold through a decision boundary. Therefore, for the clinical inference task with the provided dataset, both generative and discriminative models can be useful and demand empirical evaluation.

Table 3 also demonstrates that knowledge augmentation by adding evidence from all the sections does not improve the performance of the models in almost all cases. One possible reason is that adding more information makes it more challenging for the models to extract the relevant information. Also, by increasing the input length, the model struggles with high-dimensional input space.

Among the models, only DeBERTa-lg and FlanT5 are general-purpose models while the rest are tailored for the clinical domains by pre-training the models on domain-specific data. Also, DeBERTa-lg and SciFive are the only task-specific NLI models studied in this work. This is intriguing to observe that although DeBERTa-lg is not pre-trained on clinical data, it yields the highest F1-score. Thus, a model tailored to the task but not initially trained on domain-specific data may outperform a domain-specific model that lacks task specificity, demonstrating the importance of task-oriented adaptation rather than relying solely on domain-specific pre-training. This outcome contradicts our initial hypothesis that domain-specific pre-trained LLMs are necessary for superior performance.

Finally, the number of trainable parameters of the discriminative models is not found to be linked with model performance since discriminative BioLinkBERT, containing 340M parameters, performs either on par with or subpar than ClinicalBERT and BioClinicalBERT with 110M parameters each. However, the generative SciFive model, consisting of a larger number of trainable parameters than FlanT5, exhibits better performance in certain metrics e.g. faithfulness and consistency.

6 Conclusion

In this paper, we describe our system for the SemEval-2024 Task 2, dealing with NLI for clinical trials. Leveraging DeBERTa-lg, a discriminative pre-trained model tailored to the NLI task, we achieve a consistency score of 0.76, securing the 4th position out of 31 participants. Our exploration yields intriguing insights: both discriminative and generative models exhibit promise for this clinical inference task. In addition, we find that knowledge augmentation poses challenges for the model, possibly due to the higher dimensionality of the input space. Moreover, task-specific but not domain-specific models are found to be better performing than domain-specific but not task-specific models. However, it is worthwhile to mention that all the models are fine-tuned with the same clinical data. Interestingly, while the performance of the discriminative model is not affected by the number of parameters, it appears to influence the performance of generative models.

As part of future work, we intend to explore the applicability of parameter-efficient techniques including adapter-tuning (Houlsby et al., 2019), LoRA (Hu et al., 2021), etc. by deploying them for the clinical inference task.

7 Acknowledgments

The authors would like to thank the committee of SemEval2024, the organizers of Task 2, and the reviewers. Special thanks to Md Zobaer Hossain for providing continuous support.

References

Nancy E Avis, Kevin W Smith, Carol L Link, Gabriel N Hortobagyi, and Edgardo Rivera. 2006. Factors associated with participation in breast cancer treat-

ment clinical trials. *Journal of Clinical Oncology*, 24(12):1860–1867.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, 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.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

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 preprint arXiv:2106.09685*.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

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.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *International conference on machine learning*, pages 6905–6916. PMLR.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1863–1870, Toronto, Canada. Association for Computational Linguistics.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.