NYCU-NLP at SemEval-2024 Task 2: Aggregating Large Language Models in Biomedical Natural Language Inference for Clinical Trials

Lung-Hao Lee[1,2], **Chen-Ya Chiou**[2] **and Tzu-Mi Lin**[1]
[1]Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University
[2]Department of Electrical Engineering, National Central University
{lhlee, ltmdegf4.ii12}@nycu.edu.tw, 109501528@cc.ncu.edu.tw

Abstract

This study describes the model design of the NYCU-NLP system for the SemEval-2024 Task 2 that focuses on natural language inference for clinical trials. We aggregate several large language models to determine the inference relation (i.e., entailment or contradiction) between clinical trial reports and statements that may be manipulated with designed interventions to investigate the faithfulness and consistency of the developed models. First, we use ChatGPT v3.5 to augment original statements in training data and then fine-tune the SOLAR model with all augmented data. During the testing inference phase, we fine-tune the OpenChat model to reduce the influence of interventions and fed a cleaned statement into the fine-tuned SOLAR model for label prediction. Our submission produced a faithfulness score of 0.9236, ranking second of 32 participating teams, and ranked first for consistency with a score of 0.8092.

1 Introduction

Biomedical Natural Language Inference (NLI) seeks to determine whether a proposed statement is entailment, contradiction, or neutral according to a given clinical trial. The MEDIQA-2019 shared task (Ben Abacha et al., 2019) covered an NLI subtask in the medical domain, including clinical sentences from the MIMIC-III database (Romanov and Shivade, 2018). In this shared task, most systems were built on the BERT model (Devlin et al., 2019) and MT-DNN (Liu et al., 2019). The BERT-BiLSTM-Attention model (Lee et al., 2019) was proposed for medical text inference. The DoubleTransfer model (Xu et al., 2019) was presented to use a multi-source transfer learning approach to acquire knowledge from MT-DNN and Sci-BERT (Beltagy et al., 2019). In addition, since the evaluation data is sourced from the clinical domain, variations of BERT such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2020) were used frequently.

SemEval-2023 Task 7 (Jullien et al., 2023b) (called NLI4CT) focused on multi-evidence natural language inference for Clinical Trial Reports (CTR) (Jullien et al., 2023a). Participants should determine the inference relation (i.e., entailment or contradiction) between CTR-statements in the NLI subtask. The sentence-level and token-level encodings were exploited in a multi-granularity inference network (MGNet) (Zhou et al., 2023). The DeBERTa-v3 model (He et al., 2023) was fine-tuned on the prompted input sentences to discriminate the inference relation between the statement and clinical trials (Wang et al., 2023b). The BioLinkBERT transformer (Yasunaga et al., 2022) was used with a soft voting ensemble mechanism to enhance the NLI performance (Chen et al., 2023). The Flan-T5 model (Chung et al., 2022) was fine-tuned with instructions to explore its capabilities for multi-evidence NLI (Kanakarajan and Sankarasubbu, 2023).

Following the success of the NLI4CT-2023 task, SemEval-2024 Task 2 (Jullien et al., 2024) re-grounds this task in interventional and causal analyses of NLI models (Yu et al., 2022), with a contrast set containing the designed interventions and expected labels to investigate the faithfulness and consistency of the developed models. This task is based on the same collection of breast cancer CTRs (Jullien et al., 2023a). The statements in the training set are identical to those in the previous task, but perform a variety of interventions to statements on the development and test sets, making claims about a single CTR or comparing

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1455–1462

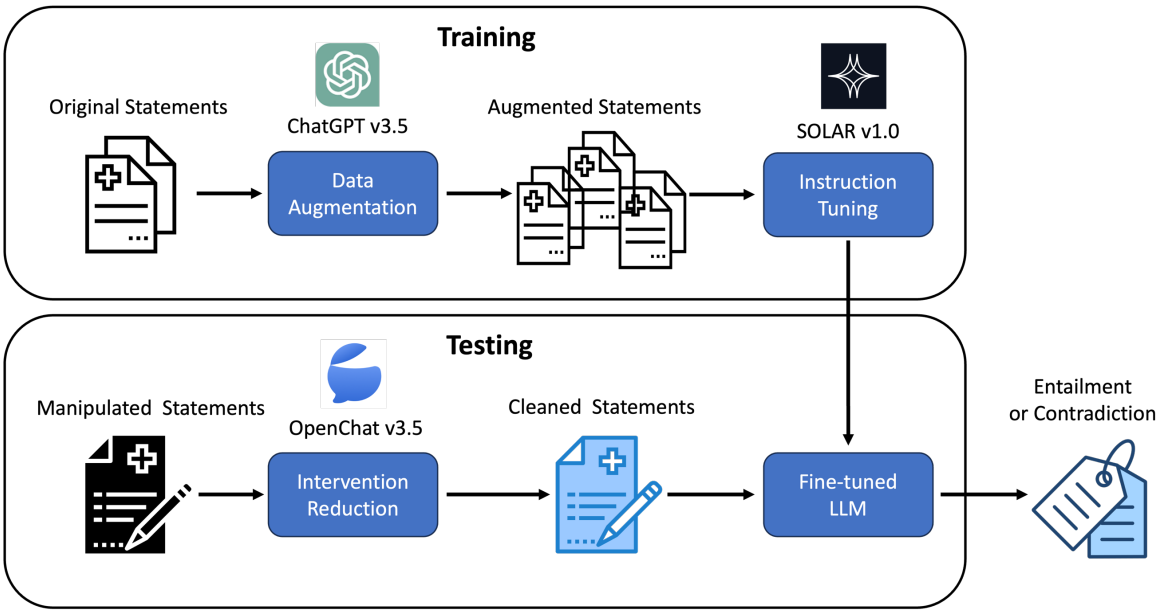

Figure 1: Our NYCU-NLP system architecture for the NLI4CT-2024 task.

two CTRs while either preserving or inversing the entailment relations. For the NLI4CT-2024 task, given a statement with/without interventions, the participating system should determine the inference relation as either entailment or contradiction.

This paper describes the **NYCU-NLP** (**N**ational **Y**ang Ming **C**hiao Tung **U**niversity, **N**atural **L**anguage **P**rocessing Lab) system for the NLI4CT-2024 task. Given the promising results obtained by Large Language Models (LLM) for various NLP tasks, we aggregate several LLMs in biomedical NLI for clinical trials. We use ChatGPT (OpenAI, 2023) to augment original statements and then fine-tune the SOLAR model (Kim et al., 2023) with instructions designed for the NLI task. Since a statement may be manipulated during testing inference phase, we first fine-tune the OpenChat model (Wang et al., 2023a) to reduce the influence of interventions. Finally, a cleaned statement along with CTRs is fed into the fine-tuned SOLAR model for label prediction (i.e., entailment or contradiction). Evaluation results show that our proposed NYCU-NLP system had a faithfulness score of 0.9236, ranking second among 32 participating teams, and ranked first for consistency with a score of 0.8092.

The rest of this paper is organized as follows. Section 2 describes the NYCU-NLP system for the NLI4CT-2024 task. Section 3 presents the results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NYCU-NLP System

Fig. 1 shows our NYCU-NLP system architecture for the NLI4CT-2024 task. Our system is composed of four main parts: 1) ChatGPT (OpenAI, 2023) for data augmentation; 2) Instruction tuning on SOLAR (Kim et al., 2023); 3) OpenChat (Wang et al., 2023a) for intervention reduction; and 4) Fine-tuned LLM for label prediction.

2.1 Data Augmentation

We use ChatGPT (OpenAI, 2023) to augment the training data for intervention adaptation. Fig. 2 shows the prompts inputted to the ChatGPT API (gpt-3.5-turbo-1106) and example outputs. We provide a system prompt to set up ChatGPT as a writer, skilled in rewriting sentences. For the first prompt, we obtain three rewritten statements without any restrictions. For the second prompt, we ask the LLM to rephrase four statements, each independently fulfilling the following requirements: 1) change as many words as possible except the existing words in primary and secondary trials; 2) change the order of existing words in the statement; 3) change percentage numbers to decimals and vice versa; and 4) change the numbers in terms of percentages and decimals to a

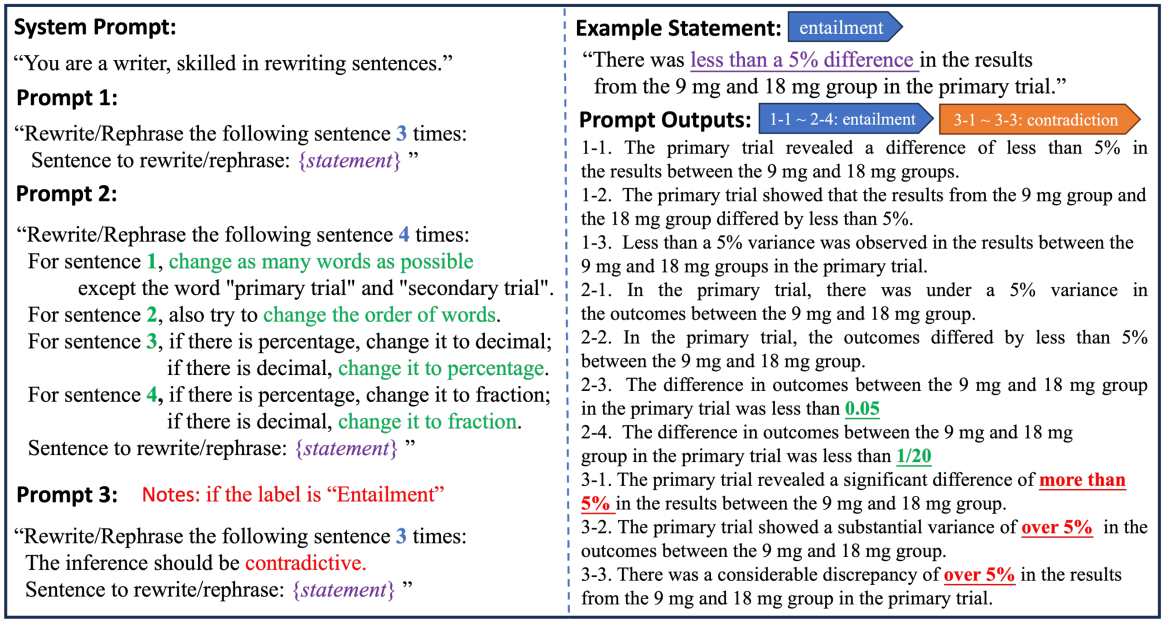

Figure 2: Prompts designed for data augmentation in ChatGPT v3.5.

fraction form. The above prompts are used for both entailment and contradiction labels. However, the third prompt is designed for the entailment label only, rewriting the original statements with contrary meanings to obtain contradictive inferences.

We also clean augmented statements to remove potentially inappropriate statements. For prompts 2-3 and 2-4, if the original statements do not contain numbers, but augmented statements contain numbers in any forms, we remove those augmented statements because these numbers are mostly hallucinations.

2.2 Instruction Tuning

We use original and augmented statements with the corresponding labels to fine-tune the SOLAR model (Kim et al., 2023). SOLAR-10.7B presents a depth up-scaling (DUS) technique to integrate Mistral 7B (Jiang et al., 2023) weights into the upscaled layers, and performs continued pre-training for the entire model. Supervised fine-tuning (SFT) and direct preference optimization (DPO) (Rafailov et al., 2023) were then used to fine-tune the model with designed instructions.

We continually fine-tune the SOLAR-10.7B-Instruct-v1.0 LLM. We use instruction tuning (Wei et al., 2022) and LoRA (Hu et al., 2021) techniques with prompts shown in Fig. 3 to optimize the SOLAR model for this NLI task. Flash attention

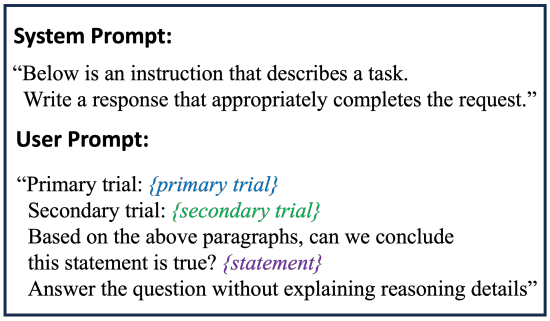

Figure 3: Prompts used for instruction tuning

(Dao et al., 2022) is also used to reduce the GPU requirements and accelerate the model fine-tuning process.

2.3 Intervention Reduction

A testing statement may be manipulated with some interventions, including numerical reasoning, vocabulary and syntax, and semantics, to investigate the consistency and faithfulness of the developed models. Technical details used to perform the interventions were not disclosed during the evaluation phase.

Therefore, we fine-tuned OpenChat v3.5 (Wang et al., 2023a) to reduce the influence of interventions. OpenChat is a framework used to advance open-source language models with mixed-quality data. As shown in Fig. 4, we used two exemplars for two-shot prompt learning. First, we

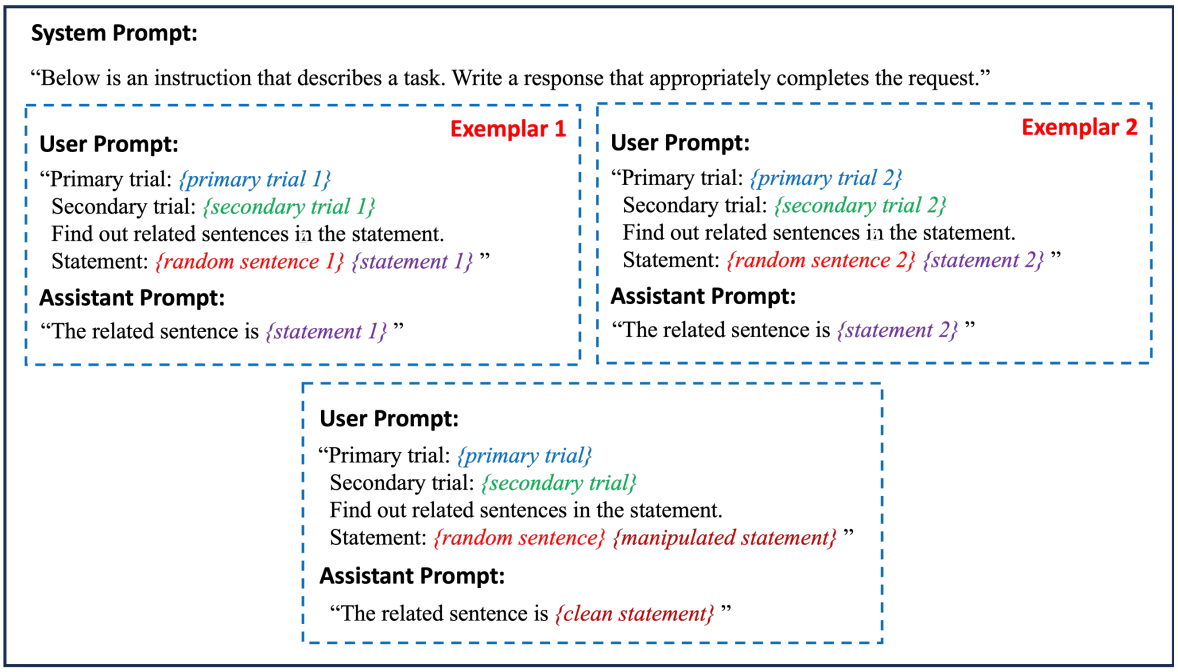

Figure 4: Prompts designed for intervention reduction in OpenChat v3.5.

randomly collected 10,000 abstracts published from Jan. 1st to Jan. 10th, 2024 from the arXiv preprint server. These were segmented into a total of 21,135 sentences. Finally, we randomly selected one sentence as an intervention sentence for both exemplars. The statements were selected from the training set, in which the first statement contains only one sentence and the second statement contains at least two sentences.

During the evaluation phase, an original statement is regarded as a manipulated statement and the fine-tuned LLM is expected to identify a cleaned statement. In most cases, a cleaned statement is a part of an original statement for intervention reduction. If an output statement contains sentences that don't belong to the original statement, the cleaned statement will be discarded and the original statement is used as the input for inference testing.

2.4 Fine-tuned LLMs for Label Prediction

Following the instruction shown in Fig. 3, the fine-tuned LLM processes a given statement based on the CTRs and answers the question without explaining its reasoning in detail. In the LLM response, if the first token is Yes or True, the predicted label is entailment, and otherwise contradiction. If the first token belongs to neither of these characteristics, we will check the vocabulary table to determine the corresponding probabilities of Entailment and Contradiction tokens. If the former exceeds the latter, the predicted label is returned as entailment, and otherwise contradiction.

3 Experiments and Results

3.1 Data

The datasets were mainly provided by task organizers (Jullien et al., 2024). A total of 1000 collected breast cancer CTRs were used as known premises. The training set used 1,700 statements to make claims about a single CTR or to compare two CTRs labelled as either entailment or contradiction. We used these statements for data augmentation, producing a total of 13,484 generated statements for LLM fine-tuning.

During the system development and evaluation phases, task organizers performed a variety of interventions on the statements in the development and test sets, either preserving or inversing the entailment relations. A total of 2,142 statements were used to develop the system and obtain the optimized parameters. Finally, the test set containing 5,500 statements was used to evaluate the system performance.

3.2 Settings

In addition to our fine-tuned SOLAR model (Kim et al., 2023), we used Mistral (Jiang et al., 2023),

Model (#para)	Development			Test		
	F1	Faithfulness	Consistency	F1	Faithfulness	Consistency
Orca2 (13B)	0.8223	0.8899	0.7914	0.7747	0.8692	0.7643
Qwen (14B)	0.8367	0.8542	0.8076	0.7657	0.8681	0.7730
Mistral (7B)	0.8500	0.9196	0.8213	0.7623	0.8611	0.7805
SOLAR (10.7B)	**0.8842**	**0.9554**	**0.8506**	**0.7790**	**0.9236**	**0.8092**

Table 1: Fine-tuned LLM results for the development and test sets.

Orca2 (Mitra et al., 2023) and Qwen (Bai et al., 2023) LLMs for performance comparison. All models were downloaded from HuggingFace[1]. We continuously fine-tuned these models using the augmented training set. All models were configured to obtain the highest average faithfulness and consistency scores on the development set. The hyperparameter values of our used SOLAR LLM were finally optimized as follows: epochs 20; batch size 8; optimizer Adafactor; learning rate schedule used a cosine decay with optional warmup; warmup ratio 0.05; max learning rate 7.5e-5; LoRA r 16; LoRA alpha 16; LoRA drop 0.05; max token length 2048 and original statement sample ratio 0.3.

3.3 Metrics

The *control F1* measures fundamental model performance of those testing instances without interventions, identical to the previous NLI4CT-2023 task and thus facilitating a direct performance comparison.

Faithfulness is estimated to measure the model's ability to correctly change its predictions when exposed to a semantic-altering intervention. The better system is expected to make the correct prediction for the correct reason.

Consistency measures the model's ability to predict the same label for original statements and contrast statements for semantic-preserving interventions. The better system is expected to produce the same outputs for semantically equivalent problems.

3.4 Results

Table 1 shows our submissions obtained consistent results for the development and test sets. The SOLAR model (Kim et al., 2023) outperformed Orca2 (Mitra et al., 2023), Quwen (Bai et al., 2023) and Mistral (Jiang et al., 2023) LLMs for all metrics.

Our SOLAR LLM achieved a control F1 score of 0.7790, significantly outperforming our submission for the NLI4CT-2023 task (F1 of 0.7091) based on ensemble BioLinkBERT transformers (Chen et al., 2023). This confirms that using LLMs properly can outperform pre-trained language models for the same task. In addition, the number of parameters in the LLM doesn't directly influence performance, indicating that model architecture is more important rather than scale.

In our proposed system workflow, regardless of which LLM model was used as the main framework for the NLI task, a higher faithfulness score was achieved when compared with the consistency score. This indicates that an LLM usually makes correct predictions with correct reasons.

In summary, in the NLI4CT-2024 task, our system based on the SOLAR model produced a promising faithfulness score of 0.9236, ranking second place among 32 participating systems, and ranked first for consistency with a score of 0.8092.

4 Conclusions

This study describes the NYCU-NLP submission for the SemEval-2024 NLI4CT task, including system design, implementation and evaluation. We aggregated several LLMs to determine the inference relation between CTRs and statements that may be manipulated with designed interventions to investigate the faithfulness and consistency of the developed models. Our system obtained a faithfulness score of 0.9236, ranking second among all 32 participating teams, and ranked first for consistency with a score of 0.8092.

[1] https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0
https://huggingface.co/microsoft/Orca-2-13b
https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca
https://huggingface.co/Qwen/Qwen-14B-Chat

Acknowledgments

This study is partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 111-2628-E-A49-029-MY3.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fanm Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Lin, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv:2309.16609v1*. https://doi.org/10.48550/arXiv.2309.16609

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pages 3615-3620. http://dx.doi.org/10.18653/v1/D19-1371

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In *Proceedings of the 18th Biomedical Natural Language Processing Workshop and Shared Task*, Association for Computational Linguistics, pages 370–379. http://dx.doi.org/10.18653/v1/W19-5039

Chao-Yi Chen, Kao-Yuan Tien, Yuan-Hao Cheng, and Lung-Hao Lee. 2023. NCUEE-NLP at SemEval-2023 Task 7: Ensemble biomedical LinkBERT transformers in multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 776-781. https://doi.org/10.18653/v1/2023.semeval-1.107

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, Yangping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quo V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416v5*. https://doi.org/10.48550/arXiv.2210.11416

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient attention with IO-awareness. *arXiv:2205.14135v2*. https://doi.org/10.48550/arXiv.2205.14135

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pages 4171-4186. http://dx.doi.org/10.18653/v1/N19-1423

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations*.

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:2106.09685v2*. https://doi.org/10.48550/arXiv.2106.09685

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv: 1904.05342v3*. https://doi.org/10.48550/arXiv.1904.05342

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee, Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv: 2310.06825v1*. https://doi.org/10.48550/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*. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donald Landers, and André 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. Association for Computational Linguistics, pages 16745-16764. https://doi.org/10.18653/v1/2023.emnlp-main.1041

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donald 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*. Association for Computational Linguistics, pages 2216-2226. https://doi.org/10.18653/v1/2023.semeval-1.307

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv: 2305.18290v2*. https://doi.org/10.48550/arXiv.2305.18290

Kamal Raj Kanakarajan, and Malaikannan Sankarasubbu. 2023. Saama AI research at SemEval-2023 Task 7: Exploring the capabilities of Flan-T5 for multi-evidence natural language inference in clinical trail data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 995-1003. https://doi.org/10.18653/v1/2023.semeval-1.137

Dahyun Kim, Chanjun Park, Sanghoom Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjim Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling. *arXiv:2312.15166v2*. https://doi.org/10.48550/arXiv.2312.15166

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. https://doi.org/10.1093/bioinformatics/btz682

Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kou-Kai Shyu. 2019. NCUEE at MEDIQA 2019: medical text inference using ensemble BERT-BiLSTM-Attention model. In *Proceedings of the 18th Biomedical Natural Language Processing Workshop and Shared Task*. Association for Computational Linguistics, pages 528-532. http://dx.doi.org/10.18653/v1/W19-5058

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 4487-4496. https://doi.org/10.18653/v1/P19-1441

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi. Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *arXiv:2311.11045v2*. https://doi.org/10.48550/arXiv.2311.11045

OpenAI. 2023. ChatGPT (Large language model). https://chat.openai.com/chat

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*. Association for Computational Linguistics, pages 1586-1596. https://www.aclweb.org/anthology/D18-1187

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. OpenChat: Advancing open-source language models with mixed-quality data. *arXiv:2309.11235v1*. https://doi.org/10.48550/arXiv.2309.11235

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023b. KnowComp at SemEval-2023 Task 7: Fine-tuning pre-trained language models for clinical trial entailment identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 1-9. https://doi.org/10.18653/v1/2023.semeval-1.1

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations*. https://openreview.net/forum?id=gEZrGCozdqR

Yichong Xu, Xiaodong Liu, Chunyuan Li, Hoifung Poon and Jianfeng Gao. 2019. DoubleTransfer at MEDIQA 2019: Multi-source transfer learning for natural language understanding in the medical domain. In *Proceedings of the 18th Biomedical Natural Language Processing Workshop and Shared Task*. Association for Computational Linguistics, pages 399-405. https://doi.org/10.18653/v1/W19-5042

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)*. Association

for Computational Linguistics, pages 8003-8016. http://dx.doi.org/10.18653/v1/2022.acl-long.551

Sicheng Yu, Jing Jiang, Hao Zhang, Yulei Niu, Qianru Sun, and Lidong Bing. 2022. Interventional training for out-of-distribution natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11627-11638. https://doi.org/10.18653/v1/2022.emnlp-main.799

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. THiFLY research at SemEval-2023 Task 7: A multi-granularity system for CTR-based textual entailment and evidence retrieval. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 1681-1690. https://doi.org/10.18653/v1/2023.semeval-1.234