iML at SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials with LLM Based Ensemble Inferencing

Abbas Akkasi[1], Adnan Khan[1], Mai A. Shaaban[2], Majid Komeili[1], Mohammad Yaqub[2],

[1]School of Computer Science Carleton University, Ottawa, Canada
[2] Mohamed bin Zayed University of Artificial Intelligence Abu Dhabi, UAE
Correspondence: abbasakkasi@cunet.carleton.ca

Abstract

The task of textual entailment holds significant importance when dealing with clinical data, as it serves as a foundational component for extracting and synthesizing medical information from vast amounts of unstructured text.

To investigate the consistency with which Natural Language Inference (NLI) models capture semantic phenomena critical for intricate inference within clinical NLI contexts, SemEval−2024 has organized a shared task focused on NLI for Clinical Trials (NLI4CT). This task provides participants with a dataset annotated by humans for the purpose of model training and requires the submission of the results on test data for evaluation. We engaged in this shared task2 at SemEval−2024, employing a diverse set of solutions, with a particular emphasis on leveraging a Large Language Model (LLM) based zero-shot inference approach to address the challenge.

1 Introduction

Clinical NLI is a specialized application of Natural Language Processing (NLP) that focuses on understanding and inferring information from text within the healthcare domain. It involves analyzing and drawing conclusions from clinical narratives, such as electronic health records (EHRs), doctor's notes, medical transcripts, clinical trials and other forms of medical documentation (Percha et al., 2022). The goal of clinical NLI is to determine the logical relationship between premises and hypotheses (conclusions) in clinical text. By inferring information from clinical text, NLI can assist healthcare providers in making informed decisions by providing evidence-based recommendations and alerts. In addition, clinical NLI can be used to identify patient cohorts for clinical trials or research studies by inferring patient eligibility based on inclusion and exclusion criteria mentioned in clinical records. Applications of clinical NLI are not limited to the ones mentioned and there are lots of other usages in which clinical NLI can be useful (Percha et al., 2021). NLI for clinical trials faces unique challenges due to the complexity of medical language, the need for domain-specific knowledge, and the sensitivity and privacy concerns associated with health data. However, advancements in NLP and specifically Large Language Models (LLMs) are continuously improving the accuracy and applicability of clinical NLI, making it an increasingly valuable tool in the healthcare industry.

To foster collaboration and dissemination of novel insights within this field, SemEval 2024 (Jullien et al., 2024) has established a shared task exclusively devoted to clinical NLI. A publicly accessible dataset, annotated by humans, has been made available to facilitate the comparison of solutions proposed by different researchers.

To address the challenge, we developed an ensemble-oriented solution that combines various Large Language Models (LLMs) based models within the framework of prompting and fine-tuned classification. Our primary goals were to first understand the comparative performance of generative models versus classification models. Subsequently, we explored whether the use of automatic summarization models to condense the premises would influence the efficacy of both classifiers and generative models. Ultimately, our approach sought to facilitate synergistic interactions among the different models, leveraging their respective strengths to mitigate individual inference limitations.

Nevertheless, despite conducting a variety of experiments that involved combining summarization, fine-tuning classifiers, prompting, and more, the results demonstrated a clear superiority of generative models in comparison to the others, even when used independently.

The remainder of this paper is organized as follows: Section 2 provides a brief review of related

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 170–174

work. The proposed model and its constituent modules are detailed in Section 3. Sections 4 and 5 discuss the experiments conducted and the corresponding results. Finally, we conclude the paper in Section 6.

2 Past Work

Recent literature underscores the need for sophisticated models that can accurately capture the semantics of clinical narratives and support reasoning in line with medical knowledge. Jullien et al. (2023), introduced a shared task on NLI for clinical trials (NLI4CT), providing a dataset of annotated clinical trials and inviting researchers to develop models to tackle the associated challenges. The shared task comprises two sub-tasks: Textual Entailment and Evidence Retrieval, each designed to advance the state of NLI systems within the clinical domain.

Zhou et al. (2023), took part in the NLI4CT-2023 challenge, proposing a model that utilizes both sentence-level and token-level encoding to address the task at hand. Furthermore, they enhanced the model's overall performance by employing general (T5-based model) and domain-specific (SciFive) pre-trained LLMs.

Kanakarajan and Sankarasubbu (2023), conducted an evaluation of several instruction-tuned Large Language Models (LLMs) in a zero-shot setting and fine-tuned the best-performing instruction-tuned model (T5 family models). Their findings suggest that instruction-tuned models yield better results for datasets with limited training samples. Additionally, they explored the impact of various prompts on the overall performance of the model. (Vladika and Matthes, 2023) and (Chen et al., 2023), both created a model based on an ensemble approach that combines various fine-tuned iterations of biomedical LLMs. These models are designed to extract evidence from clinical trial report premises to support textual entailment in specific statements. Wang et al. (2023), developed a system that utilizes prompts created by humans to gather information from statements, section titles, and clinical trials. They then fine-tune pre-trained language models on these prompted sentences, training the models to identify the inferential connections between the statements and the clinical trials. Pahwa and Pahwa (2023), characterized the NLI task as a form of text pair classification and utilized the GPT-3 model to classify samples within the framework of few-shot prompting. This approach takes advantage of the semantic similarity between text samples and the examples provided for in-context learning.

Dias et al. (2023), employed supervised contrastive learning to enhance the sentence pair representations in the Biomed RoBERTa model. They then fine-tuned a linear classifier built upon these improved representations to identify evidence and execute textual entailment classification for sentence pairs.

Vassileva et al. (2023), introduced a two-tiered system to address the sub-tasks of NLI4CT-2023. Initially, the system employs a BERT-based classifier, supplemented by contextual data augmentation, to categorize evidence-statement pairs as relevant or irrelevant. Subsequently, leveraging the relevant segments of the clinical trial identified in the first stage, the system applies another BERT-based classifier to ascertain whether the relationship between the elements is one of entailment or contradiction.

Volosincu et al. (2023), illustrated that a transformer model pre-trained on biomedical data for the task of entailment relation in NLI4CT-2023 does not automatically outperform traditional approaches like CNNs. Nonetheless, their model exceeded the baseline system's performance and provided meaningful directions for future research on how the model's architecture can be developed further.

3 Proposed Model

In tackling the NLI4CT task, our approach involved the construction of an ensemble model that integrates the judgments of multiple distinct decision-makers. These decision-makers differ concerning the nature of input data they process, the foundational models they employ, and the methodologies they adopt for label determination. Figure 1 provides a comprehensive illustration of the proposed solution. Components of the ensemble pool were developed within the frameworks of classification or prompting, utilizing LLMs. For classification tasks, SciFive (Zhou et al., 2023) was selected as the base model due to its exemplary performance in the NLI4CT-2023 task. To enhance the models' ability to assimilate information from the input data, we employed both extractive and abstractive summarization techniques. The abstractive summarization was conducted using the T5-large model (Raffel et al., 2020) to condense the premises. For

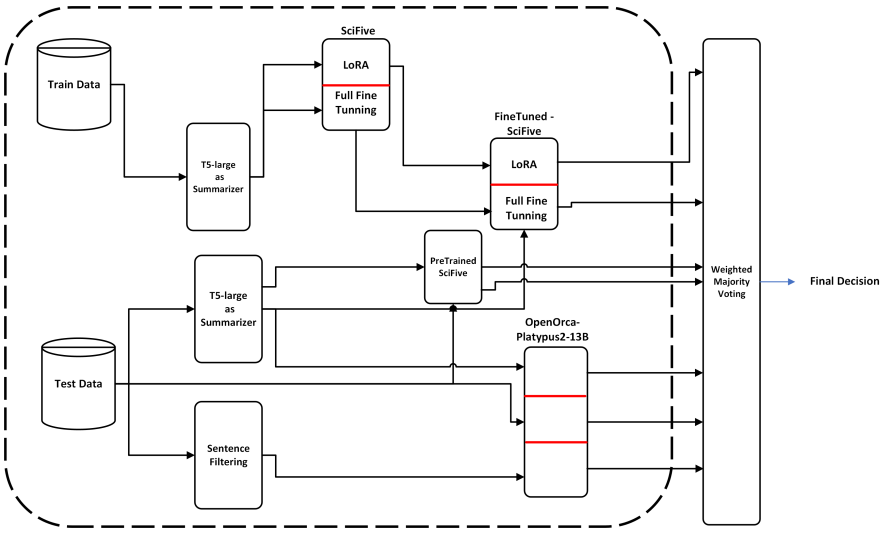

Figure 1: Ensemble Model Proposed

extractive summarization, the premises were initially segmented into individual sentences, after which those exhibiting lower semantic similarity to the hypothesis were excluded.

The pre-trained SciFive model ingests the text summarized by T5 to generate the initial component of the ensemble pool. Subsequently, this model undergoes fine-tuning through two distinct methodologies utilizing the summarized data: comprehensive fine-tuning and parameter-efficient fine-tuning, the latter of which is facilitated by employing LoRA (Hu et al., 2021) to produce subsequent members of the ensemble pool.

The remaining decision-makers within the ensemble are derived by prompting generative LLM[1] in a zero-shot inference context, utilizing both the original input data and variously summarized inputs. The specific prompt employed for the model is delineated in Listing 1.

```
# For Type="Comparison"
prompt = f''' Assess the logical
    relationship between two clinical
    trial descriptions (Primary Trial (
    PT), Secondary Trial: (ST)) as
    premises and the hypothesis given
    below.
     Return 'Entailment' if the premises
     logically imply the hypothesis, and
     'Contradiction' if the hypothesis
     conflicts with the information in
     the premises.
     Primary Trial (PT) : {PE}
     Secondary Trial (ST): {SE}
     hypothesis: {hypothesis}
'''
# For Type="Single"
prompt = f'''Evaluate the logical
    relationship between the clinical
    trial premise (PE) and the
    hypothesis given below.
     Return 'Entailment' if the premise
    logically implies the hypothesis,
    and 'Contradiction' if the
    hypothesis conflicts with the
    information in the premise.
Clinical Trial (PE): {PE}
hypothesis: {hypothesis}
'''
```

Listing 1: Prompt Template Used.

Ultimately, the final decision for the test samples were made using a weighted majority voting approach. The performance of models on *practice_test* set were used for the combination process.

4 Experiments

We have conducted our experiments utilizing the dataset provided by the task's organizers that is explained in Section 4.1. For the models based on prompting, we utilized only the test and practice_test datasets, whereas the training data was employed exclusively for fine-tuning the classification-based models. Beyond experimenting with models within our ensemble framework, we also explored the integration of results from fine-tuned classification models as a form of external knowledge within the context of prompting. The efficacy of all models is evaluated using three metrics: Macro

[1]OpenOrca-Platypus2-13B, which is an autoregressive language model that utilizes the Lllama 2 transformer architecture. It is tailored for a variety of general-use applications, including chat, text generation, and code generation. This model has undergone training with a diverse mix of datasets, focusing on STEM and logic-based content, and it incorporates a carefully selected portion of data from the GPT-4 dataset within the OpenOrca collection.

F1 Score, Faithfulness, and Consistency, each of which is briefly described in Section 4.2.

4.1 Dataset

The corpus presented for analysis encompasses training, development, practice_test, and test datasets, each containing a distinct number of samples. Table 1 displays the quantity of samples for each dataset. The content of each sample, including statements and evidence, has been reconstructed by a collaborative effort of clinical domain experts, clinical trial organizers, and research oncologists associated with the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team.

Split	#Samples	#Entailment	#Contradiction
Train	1700	850	850
Practice_test	2142	730	1412
Development	200	100	100
Test	5500	1841	3659

Table 1: Overview of Dataset Splits: Distribution of Samples, Entailment, and Contradiction Labels

4.2 Evaluation

In assessing system performance, the organizers, in conjunction with the macro F1 score, opted to examine model efficacy on a contrast dataset comprising statements with interventions. The comprehensive ranking of the systems is determined by the mean of two novel metrics: Faithfulness (as defined in Equation. 1) and Consistency (as defined in Equation. 2), across all types of interventions.

$$FaithFulness = \frac{1}{N}\sum_{1}^{N} |f(y_i) - f(x_i)| \quad (1)$$

where $x_i \in C : Label(x_i) \neq Label(y_i), f(y_i) = Label(y_i)$.

$$Consistency = \frac{1}{N}\sum_{1}^{N} 1 - |f(y_i) - f(x_i)| \quad (2)$$

where $x_i \in C : Label(x_i) = Label(y_i)$. Faithfulness quantifies the degree to which a system reaches an accurate prediction based on the correct rationale. While, Consistency measures the degree to which a system yields identical outputs for semantically equivalent queries. The results obtained during the experimental trials are presented in the subsequent section.

5 Results

The performance result of individual models within the ensemble, as applied on both practice_test and test datasets, are illustrated in Table 2.

The proposed model exhibits faithfulness and consistency scores of 28% and 52%, respectively, suggesting a necessity for more robust models to effectively manage clinical trials involving diverse data types. The findings reveal that the proposed overall model performs similarly to the generative model in the prompting context. This similarity underscores the considerable potential of generative LLMs. These models can achieve better performance when instruction tuning is applied with domain-specific data. Additionally, using classification results as external knowledge for the prompting model showed minimal impact. Moreover, the use of extractive summarization yielded the lowest results, aligning with our expectations. This approach, which focuses on the similarity between individual sentences and the statement, can lead to a loss of comprehension of the entirety of the premises.

6 Conclusion

In conclusion, our participation in NLI4CT-2024 involved proposing an ensemble approach that incorporated multiple decision-makers, with two Large Language Models (LLMs) serving as foundational models. We explored various data preparation techniques, including abstractive summarization and similarity-based sentence filtering, for use in both prompting and classification contexts. The comparable performance of the prompt-based model to the overall ensemble model, coupled with its significant outperformance of the classification models, underscores the substantial potential of pretrained generative foundation models in solving similar problems. We posit that the application of instruction tuning and the incorporation of domain-specific data could markedly enhance the results.

7 Acknowledgments

We extend our sincere gratitude to Dana Osama and Anees Hashmi for their valuable cooperation and contributions to this work. In addition, this work received support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Digital Alliance of Canada, to whom the authors extend their gratitude.

	Practice_Test							
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Score	66.66	72.64	66.89	72.65	68.12	60.66	69.12	**72.65**
	Test							
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Score	66.84	65.37	66.30	69.61	66.36	52.95	66.07	**70.27**

Table 2: Performance comparison in terms of F1-score on practice test and test Datasets: M1: Pretrained SciFive, M2: Full Fine-tuned SciFive (Summarized Data), M3: Fine-tuned SciFive (LoRA and Summarized Data), M4: Prompting, M5: Prompting with Summarized Data, M6: Prompting with Filtered Sentences, M7: SciFive Results as External Knowledge for Prompting, M8: Ensemble Method

References

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 (SemEval-2023)*, pages 776–781.

Abel Corrêa Dias, Filipe Dias, Higor Moreira, Viviane Moreira, and João Luiz Comba. 2023. Team inf-ufrgs at semeval-2023 task 7: Supervised contrastive learning for pair-level sentence classification and evidence retrieval. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 700–706.

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*.

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. 2023. 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.

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 trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003.

Bhavish Pahwa and Bhavika Pahwa. 2023. Bphigh at semeval-2023 task 7: Can fine-tuned cross-encoders outperform gpt-3.5 in nli tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944.

Bethany Percha, Kereeti Pisapati, Cynthia Gao, and Hank Schmidt. 2021. Natural language inference for clinical registry curation. *medRxiv*, pages 2021–06.

Bethany Percha, Kereeti Pisapati, Cynthia Gao, and Hank Schmidt. 2022. Natural language inference for curation of structured clinical registries from unstructured text. *Journal of the American Medical Informatics Association*, 29(1):97–108.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sylvia Vassileva, Georgi Grazhdanski, Svetla Boytcheva, and Ivan Koychev. 2023. Fmi-su at semeval-2023 task 7: Two-level entailment classification of clinical trials enhanced by contextual data augmentation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1454–1462.

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. *arXiv preprint arXiv:2304.13180*.

Mihai Volosincu, Cosmin Lupu, Diana Trandabat, and Daniela Gifu. 2023. Fii smart at semeval 2023 task7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 212–220.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023. 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 (SemEval-2023)*, pages 1–9.

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. *arXiv preprint arXiv:2306.01245*.