SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data

Maël Jullien[1], Marco Valentino[2, 1], Hannah Frost[1, 3], Paul O'Regan[3], Donal Landers[3], André Freitas[1, 2]
Department of Computer Science, University of Manchester, United Kingdom[1]
Idiap Research Institute, Switzerland[2]
Digital Experimental Cancer Medicine Team, Cancer Research UK Manchester Institute[3]
{firstname.surname}@manchester.ac.uk
{Paul.ORegan,Donal.Landers}@digitalecmt.org

Abstract

This paper describes the results of SemEval 2023 task 7 – Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT) – consisting of 2 tasks, a Natural Language Inference (NLI) task, and an evidence selection task on clinical trial data. The proposed challenges require multi-hop biomedical and numerical reasoning, which are of significant importance to the development of systems capable of large-scale interpretation and retrieval of medical evidence, to provide personalized evidence-based care.

Task 1, the entailment task, received 643 submissions from 40 participants, and Task 2, the evidence selection task, received 364 submissions from 23 participants. The tasks are challenging, with the majority of submitted systems failing to significantly outperform the majority class baseline on the entailment task, and we observe significantly better performance on the evidence selection task than on the entailment task. Increasing the number of model parameters leads to a direct increase in performance, far more significant than the effect of biomedical pre-training. Future works could explore the limitations of large models for generalization and numerical inference, and investigate methods to augment clinical datasets to allow for more rigorous testing and to facilitate fine-tuning.

We envisage that the dataset, models, and results of this task will be useful to the biomedical NLI and evidence retrieval communities. The dataset[1], competition leaderboard[2], and website[3] are publicly available.

1 Introduction

Clinical trials are indispensable for experimental medicine as they test the efficacy and safety of novel treatments (Avis et al., 2006). Clinical Trial Reports (CTRs) are documents that detail the methodology and results of a trial, implemented to guide personalized and targeted interventions for patients. However, there are 400,000+ published CTRs, with an increasing number being published every year (Bastian et al., 2010), making it impractical to manually carry out comprehensive evaluations of all the relevant literature when designing new treatment protocols (DeYoung et al., 2020).

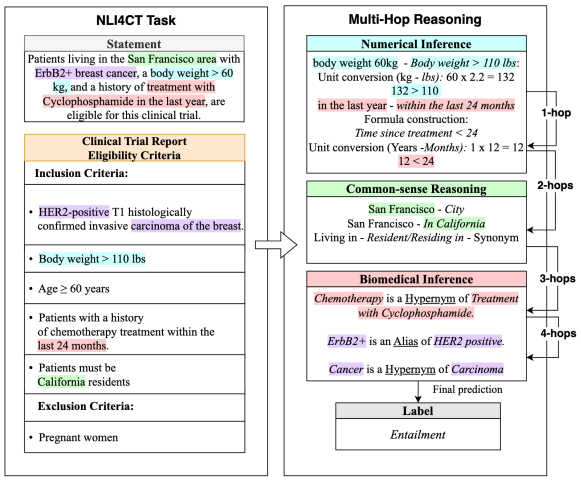

Figure 1: We propose two tasks for reasoning on clinical trial data expressed in natural language. Firstly, to predict the entailment of a **Statement** and a **CTR** premise, and secondly, to extract evidence to support the label.

To address this challenge, Natural Language Inference (NLI) (Bowman et al., 2015; Devlin et al., 2019) offers a potential solution for the large-scale interpretation and retrieval of medical evidence, to support a higher level of precision and efficiency in personalized evidence-based care (Sutton et al., 2020).

SemEval-2023 Task 7 – Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT) – is based on the NLI4CT dataset which contains 2 tasks on breast cancer CTRs, shown in

[1] https://github.com/ai-systems/nli4ct
[2] https://codalab.lisn.upsaclay.fr/competitions/8937#learn_the_details
[3] https://sites.google.com/view/nli4ct/

Figure 1. Firstly to determine the inference relation between a natural language statement, and a CTR. Secondly, to retrieve supporting facts from the CTR(s) to justify the predicted relation.

The inference task requires Multi-hop reasoning, that is the ability to combine information from multiple pieces of text to draw inferences (Jansen et al., 2018; Dalvi et al., 2021). Previous works have shown that although multi-hop reasoning can be implemented on large-scale scientific tasks, there is a significant drop-off in performance as the number of necessary hops increases (Valentino et al., 2022, 2021; Thayaparan et al., 2022, 2021). A large proportion of the NLI4CT dataset instances require the construction of inference chains in this drop-off range.

Additionally, numerical and quantitive reasoning is required to perform inference on NLI4CT, exemplified in Figure 1. Studies have shown that transformer-based models are unable to consistently apply this type of reasoning, instead relying on shallow heuristics for predictions (Patel et al., 2021; Ravichander et al., 2019; Galashov et al., 2019).

In the NLI4CT inference task, both the multi-hop and the numerical reasoning have the added complexity of being applied to CTRs. Studies have demonstrated that the word distribution shift from general domain corpora to biomedical corpora, such as CTRs, caused by the increased prevalence of aliases, acronyms, and biomedical terminology represents a significant detriment to model performance (Lee et al., 2019; Grossman Liu et al., 2021; Shickel et al., 2017; Jiang et al., 2011; Moon et al., 2015; Jimeno-Yepes et al., 2011; Pesaranghader et al., 2019; Jin et al., 2019; Wu et al., 2015).

This word distribution shift challenge is also present in the evidence selection task. Although the evidence selection task is arguably simpler than the inference task its importance cannot be understated. State-of-the-art NLI models consistently struggle to attend to relevant pieces of evidence when applied to large texts (DeYoung et al., 2020). Additionally, the ability to filter out irrelevant pieces of text reduces the likelihood of distractors (Mishra and Sachdeva, 2020) and reduces the length of the input for inference, improving efficiency.

This paper introduces SemEval-2023 Task 7 – Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT) – for biomedical NLI and evidence extraction and presents a detailed analysis of the performance of the participating systems. We report the following conclusions;

- The highest scoring system (Zhou et al., 2023) @THiFLY achieved an F1 score of 0.856 and 0.853 on the entailment task and the evidence selection task respectively.
- The tasks are challenging, most submissions did not significantly outperform the majority class baseline on the entailment task.
- On average, performance on the evidence selection task was higher than on the entailment task.
- Increasing the number of model parameters leads to a direct improvement in performance, far out-weighting the effect of biomedical pretraining.

2 Related Works

There are many existing expert-annotated resources for clinical NLI. The TREC 2021 Clinical Track (Soboroff, 2021) is a large-scale information retrieval task to match patient descriptions to clinical trials for which they are eligible. Evidence Inference 2.0 (DeYoung et al., 2020) is a Question-Answering (QA) task and span selection task, where provided with an outcome, an intervention, and a comparator intervention, systems must infer if the intervention resulted in a significant increase, significant decrease, or produced no significant difference in the outcome measurement, compared to the comparator, and identify spans that support this inference. The MEDNLI (Romanov and Shivade, 2018a) dataset is an entailment task to infer the entailment relation between a short piece of text extracted from medical history notes, and an annotated statement.

None of the aforementioned tasks encompass the full complexity of NLP over CTRs, that this the capability to reason over all sections CTRs and to simultaneously carry out biomedical and numerical inference. Instead choosing to focus on one specific CTR section. Additionally, these tasks often have repetitive inference chains, i.e. matching statements for eligibility, or retrieving measurements and comparing them. In contrast, NLI4CT instances cover all CTR sections and contain minimal repetition in inference chains, as there is no set template for statements.

Currently, Large Language Models (LLM) achieve the best results for clinical NLI (Gu et al., 2021; DeYoung et al., 2020). However, they suffer from a plethora of issues. LLMs demonstrate poor performance on quantitative reasoning and numerical operations within NLI (Ravichander et al., 2019; Galashov et al., 2019). Additionally, there is a notable drop in performance for LLMs pretrained on general domain data when applied to biomedical tasks (Lee et al., 2019), partially aggravated by a lack of well-annotated clinical data (Kelly et al., 2019). NLI4CT is designed to assist in the development and benchmarking of models for clinical NLI.

3 Task Description

NLI4CT contains two tasks, Task 1, textual entailment, and Task 2, evidence selection. Each instance in NLI4CT contains a CTR premise and a statement. Premises contain 5-500 tokens, describing either the results, eligibility criteria, intervention, or adverse event of a trial, and the statements are sentences with a length of 10-35 tokens (see example in Figure 1), which make one or more claims about the premise. On average 7.74/21.67 facts within the premise are labeled as evidence. There are two types of instances in NLI4CT; single instances where the statement makes a claim about one CTR, and comparison instances, where the statement makes claims comparing and contrasting two CTRs. To summarize:

Task 1 Classify the inference relation between a CTR premise and a statement, as either an entailment or a contradiction, as shown in Figure 1.

Task 2 Output a subset of facts from the CTR premise, necessary to justify the class predicted in Task 1.

4 Dataset

The premises in NLI4CT are obtained from 1000 publicly available English language Breast cancer CTRs published on ClinicalTrials.gov. This data is maintained by the U.S. National Library of Medicine and is subject to the HIPAA Privacy Rule. The CTRs are split into 4 sections:

- **Eligibility criteria** - A set of conditions patients must meet to participate in the trial.
- **Intervention** - Detailed description of the type, dosage, frequency, and duration of treatments being studied.
- **Results** - Reports the results of the patient cohorts in the trial with respect to a given outcome measurement.
- **Adverse Events** - Reports the (serious) signs and symptoms observed in patients during the clinical trial.

A group of domain experts, including clinical trial organizers from a major cancer research center, took part in the annotation task. The annotators were given two CTR premises to generate an entailment statement. This is a short text that makes an objectively true claim about the contents of the premise. Annotators could choose to write a statement about one, or both premises. Non-trivial statements typically involve summarization, comparison, negation, relation, inclusion, superlatives, aggregation, or rephrasing, and require understanding multiple rows of the premise. Then the annotators select a subset of facts from the premise(s) that support the claims in the statement.

Then a negative rewriting technique (Chen et al., 2019) was applied, modifying the previously produced entailment statement to contain objectively false claims while retaining the original sentence structure and length. This technique is used to reduce the likelihood of stylistic or linguistic patterns pertaining to either entailment or contradictory statements. Annotators then extract a subset of facts from the premise that contradict the claims in the false statement,

The resulting dataset includes 2400 annotated statements with labels, premises, and evidence. The dataset was split 70/20/10 train/test/dev. The two classes and four sections are evenly distributed throughout the dataset and its splits.

5 Evaluation

The same strategy is adopted for the evaluation of the results of both Tasks. Task 1, the textual entailment task, is a binary classification task, so performance is measured using Precision, Recall, and Macro F1-score, comparing predicted labels against the gold labels. We also frame Task 2, the evidence selection task, as a binary classification task, classifying each fact in the premise as either relevant evidence, or irrelevant, we compare the predicted labels against the gold labels and compute the Precision, Recall, and Macro F1-score.

Technique/Model Type	Submissions #
Generative LLMs	8
Discriminative LLMs	16
Ontology-based	1
Semantic rule-based	1
Biomedical Pre-training	12

Table 1: Summary of the techniques and models implemented in the submissions

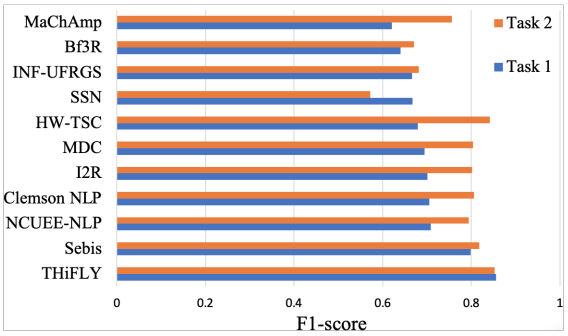

Figure 2: Graph comparing system F1 scores across the entailment task and the evidence selection task

6 Architectural Paradigms

We observe 5 different categories of approaches described in the system papers, recorded in Table 1. Generative language models are designed to learn the joint probability distribution of P(X,Y), where X is the input text, such as the statements or CTR premises, and Y is a probability output by a classification layer or a generated label from a decode-only transformer. Conversely, discriminative language models encode the conditional probability P(Y|X), designed to encode the decision boundary between different classes. Biomedical pre-training refers to the technique of training a model on a large, unlabeled biomedical dataset, such as scientific articles or patient health records. This is used to encode general features and patterns within a domain, before fine-tuning a specific task. Semantic rule-based models perform inference based on a set of human-defined asserted facts or axioms. Ontologies capture the categories, properties, and relations between the concepts of a particular domain. Ontology-based models extract entities from the input text and map them to nodes within the ontology to enrich the inputs with domain knowledge.

6.1 Transformers

The majority of submitted systems leverage discriminative transformers-based models. As shown in Table 1, 16 participants integrated discriminative transformers-based models (Vaswani et al., 2017) into their submitted systems. Generally, a task-specific output layer is appended to the pre-trained layers and fine-tuned on the training to output the probability of a statement being entailed, or a piece of evidence being relevant. Alternatively, 8/21 participants submitted systems based on generative models, as seen in Table 1. These models are either appended with a task-specific output layer and fine-tuned to output a probability or directly output entailment/contradiction or relevant/irrelevant labels.

6.2 Biomedical Pre-training

The majority of participants leverage Biomedical pre-training in their systems. As previously described LLMs trained on general domain corpora generalize poorly on biomedical corpora (Lee et al., 2019). Therefore many participants choose to apply models that are pre-trained on biomedical texts, such as datasets of scientific articles and patient health records.

7 Results and Discussion

During the 21-day evaluation period (January 10^{th}-31^{st}, 2023), 40 participants submitted a total of 643 submissions for the entailment task, and 23 participants submitted a total of 364 submissions for the evidence selection task. In total 21 participants submit system papers. Submissions for which a system paper was not provided are omitted from the tables and discussion.

The majority of systems fail to significantly outperform the majority-class baseline on the entailment task. Table 2 shows the F1 score, Recall, and Precision for Task 1. The collected results indicate that these tasks are challenging, with the majority of systems failing to achieve significantly above the majority-class baseline (0.667 F1) results on the entailment task. In particular, we observe several systems reporting 0.9-0.95 Recall, and 0.5-0.55 Precision, indicating the systems were almost exclusively predicting the "entailment" class. All systems with submitted papers significantly outperform the random baseline (0.5 F1).

The top-performing systems achieve significant gain across both tasks. Zhou et al. (2023) @THiFLY, Kanakarajan and Sankarasubbu (2023)

Work @Team name	Approach	Generative/ Discriminative	Retrieval type	Pre-training Datasets	Task 1			Task 2		
					F1	Precision	Recall	F1	Precision	Recall
(Zhou et al., 2023) @THiFLY	MGNet, BiLSTM and SciFive model ensembling	G + D	Post	PubMed Abstract, PMC	0.856	0.856	0.856	0.853	0.811	0.898
(Kanakarajan and Sankarasubbu, 2023) @Saama AI Research	Instruction-finetuned LLMs, Flan-T5	G + D	-	-	0.834	0.768	0.912	-	-	-
(Vladika and Matthes, 2023) @Sebis	Ensemble of a pipeline and joint system based on DeBERTa-v3	D	Pre	-	0.798	0.777	0.820	0.818	0.772	0.868
(Wang et al., 2023) @KnowComp	DeBERTa-v3-large.	D	-	-	0.764	0.757	0.772	-	-	-
(Chen et al., 2023) @NCUEE-NLP	Soft voting ensemble mechanism based on BioLink/BioBERT	D	Pre	MultiNLI, MedNLI, and SNLI	0.709	0.668	0.756	0.794	0.803	0.786
(Alameldin and Williamson, 2023) @Clemson NLP	GatorTron-BERT	D	Pre	UFHS notes, MIMIC-III, WikiText, PMC, and extracted CTRs	0.705	0.654	0.764	0.806	0.802	0.811
(Rajamanickam and Rajaraman, 2023) @I^2R	Evidence level inferences with T5	G + D	Pre	-	0.701	0.550	0.968	0.802	0.797	0.807
(Bevan et al., 2023) @MDC	PubMedBERT for evidence retrieval, and BioLinkBERT classifies entailment.	D	Pre	PubMed abstracts, PMC	0.695	0.668	0.724	0.804	0.814	0.795
(Zhao et al., 2023) @HW-TSC	Zero-shot ChatGPT for entailment and DeBERTaV3 for retrieval.	G + D	Post	-	0.679	0.592	0.796	0.842	0.816	0.871
(Pahwa and Pahwa, 2023) @BpHigh	Few-shot GPT-3.5 Davinci	G	-	-	0.679	0.523	0.968	-	-	-
(Feng et al., 2023) @YNU-HPCC	BioBERT, supervised contrastive learning, and back translation.	D	-	PubMed Abstracts, PMC	0.679	0.621	0.748	-	-	-
(Alissa and Abdullah, 2023) @JUST-KM	Role-based Double Roberta-Large	D	-	-	0.670	0.529	0.912	-	-	-
(Noor Mohamed and Srinivasan, 2023) @SSNSheerinKavitha	Semantic Rule based Clinical Data Analysis, TF-IDF, and BM25	-	Post	-	0.667	0.500	1.00	0.572	0.542	0.606
(Corrêa Dias et al., 2023) @INF-UFRGS	EvidenceSCL using a modified PairSCL model and pre-trained Biomed RoBERTa checkpoints.	D	Pre	Semantic Scholar corpus	0.666	0.500	0.996	0.681	0.615	0.764
(Takehana et al., 2023) @Stanford MLab	Bio+Clinical/Distil/Bio Discharge Summary BERT, and ELECTRA Small ensemble	D	-	MIMIC-III, PubMed Abstracts, PMC	0.662	0.575	0.780	-	-	-
(Conceição et al., 2023) @lasigeBioTM	Biomedical Ontology annotations, using Scispacy	-	-	-	0.661	0.511	0.936	-	-	-
(Neves, 2023) @Bf3R	Sentence-based BERT similarity model pre-trained on ClinicalBERT embeddings.	D	Post	MIMIC III	0.640	0.497	0.900	0.671	0.583	0.789
(Volosincu et al., 2023) @FII SMART	BioBERT model and a CNN model	D	-	PubMed Abstracts, PMC	0.596	0.582	0.612	-	-	-
(Vassileva et al., 2023) @FMI-SU	Contextual Data Augmentation to fine-tune BioM-BERT-Large	D	-	PubMed Abstracts, PMC, EN Wiki + Books	-	-	-	0.827	0.779	0.881
(Huang et al., 2023) @CPIC	Ensembled GPT-2 models with different parameter sizes and random seeds.	G + D	-	-	-	-	-	0.810	0.789	0.833
(Mahendra et al., 2023) @ITTC	BM25 and Word Mover Distance	-	-	-	-	-	-	0.719	0.579	0.948

Table 2: Summary of the techniques and models implemented in the leaderboard submissions. (G) Generative model, (D) Discriminative model, (Post) Evidence retrieved after entailment, (Pre) Evidence retrieved before entailment.

@Saama AI Research, Vladika and Matthes (2023) @Sebis and Wang et al. (2023) @KnowComp achieve significantly above 0.7 F1 on the entailment task. Zhou et al. (2023) @THiFLY, Zhao et al. (2023) @HW-TSC, Vassileva et al. (2023) @FMI-SU, Vladika and Matthes (2023) @Sebis, Huang et al. (2023) @CPIC, Alameldin and Williamson (2023) @Clemson NLP, Bevan et al. (2023) @MDC and Rajamanickam and Rajaraman (2023) @@I^2R surpass 0.8 F1 on the evidence selection task.

The entailment task is more challenging than the evidence selection task. Table 2 shows the F1 score, Recall, and Precision for the evidence selection task. On average systems report a +0.07 higher F1 score on the evidence selection task than on the entailment task, shown in Figure 2. This result was expected as the evidence selection task does not require systems to learn complex decision boundaries between the classes or to perform numerical inference.

Submitted systems report higher Recall than Precision. On the evidence selection task, the

vast majority of systems record a higher Recall than Precision, with an average difference of +0.055 higher Recall, this disparity is increasingly important with the top 5 systems, with an average difference of +0.077. A potential cause for the disparity between Precision and Recall results is statements such as "Patients with liver disease are eligible for the primary trial" where the full eligibility criteria must be returned, to provide evidence that there are no conditions against liver disease. This incentivizes systems to retrieve a large proportion of the premise, and perhaps more importantly to intentionally retrieve pieces of text that are not relevant to entities contained in the statement (liver disease). However, we hypothesize that the cost of incorrectly labeling relevant information as irrelevant is much more significant than the cost of including distracting information. This is because the entailment of a statement is often dependent on a single line of a premise. Therefore maximizing Recall, even at the cost of Precision may significantly improve evidence completeness.

7.1 Foundational Model Architectures

Generative models outperformed discriminative models on the entailment task. As shown in Table 2 the top 2 systems on the entailment task are based on generative models, specifically 2 variants of the T5 model (Raffel et al., 2020), SciFive (Phan et al., 2021) and Flan-T5 (Chung et al., 2022). Both of these models significantly outperform the next best system with +0.058 and +0.036 F1 respectively. It should be noted that SciFive is implemented in Zhou et al. (2023) @THiFLY, as part of an ensemble with Multi-granularity Inference Networks and BiLSTMs, and therefore the system results cannot be solely attributed to the generative components.

DeBERTa-v3 outperforms other discriminative transformer-based models on both tasks. DeBERTa-v3-based systems consistently outperform systems that apply discriminative models, on both tasks. This is also observed across a range of different systems settings (Vladika and Matthes, 2023; Zhao et al., 2023; Wang et al., 2023). DeBERTa-v3 remains competitive with the top generative approaches.

Increase in model size is correlated with an increase in performance. An increase in model size, as in models with a higher number of parameters, is strongly correlated with better performance on both Tasks. The top 5 systems in both tasks are exclusively composed of Mega Language Models (MLM) such as T5 and DeBERTa-v3-large. Additionally, Vladika and Matthes (2023); Kanakarajan and Sankarasubbu (2023) and Wang et al. (2023) all report MLMs significantly outperforming comparatively smaller models within their individual systems.

7.2 Rule-based systems

Rule-based approaches are less competitive than MLMs. Conceição et al. (2023) @lasige-BioTM experiments with a hybrid system, using the *en_core_sci_lg* spaCy pipeline to extract entities from CTR premises and retrieving their ancestors from biomedical ontologies, then computing the shortest dependency path between entities, assisted with Counts and Measurements Rules to process numerical values. The premise is then combined with the premise and classified using cosine similarity. Noor Mohamed and Srinivasan (2023) @SSNSheerinKavitha applies a semantic rule-based system consisting of a Negation equivalence rule, Double negation rule, Deductive reasoning rule, and a Condition-based equivalence rule. Classification is obtained using TF-IDF vectors and RBF-Kernel distance similarity, and evidence is selected using BM25. As seen in Table 2, these systems are not competitive with the top-performing MLMs, however, if this disparity could be corrected, symbolic models inherently offer a higher level of transparency and interpretability than current neural models.

7.3 Data augmentation

Data-augmentation does not result in a significant performance increase. Corrêa Dias et al. (2023) investigates transfer learning opportunities by adding a neutral class to NLI4CT, and merging it with MultiNLI (Williams et al., 2018) and MedNLI (Romanov and Shivade, 2018b) to train their system. Vassileva et al. (2023) @FMI-SU annotates premise facts with structural context information, attaching trial names, cohort numbers, and parent subsection headings. They observed that the trial name does not improve performance, in some cases even adding noise, but showed some improvement with cohort and subsection annotations. Alameldin and Williamson (2023) @Clemson NLP compile an additional 9000 CTRs, and train a GatorTron model with a masked-language modeling objective for one epoch, before fine-tuning

on NLI4CT. Results from this experiment reveal minor performance gains from the additional training data. Takehana et al. (2023) @Stanford Mlab uses a combination of back translation, synonym replacement, Random insertions, deletions, and swapping of words on NLI4CT to quadruple the size of the training set. The results presented in Table 2 demonstrate that data augmentation does not inherently result in improved performance, and highlight the importance of selecting suitable tasks, data, or annotations, with respect to the target domain.

7.4 Biomedical Pre-training

There is no consistently superior biomedical pre-training strategy. Models pre-trained on the PubMed Abstract and PubMed Central (PMC) datasets were implemented in 6/21 systems, including the top-performing system Zhou et al. (2023) @THiFLY. Additionally, 4/21 systems use models pre-trained on MIMIC III. There is no observable correlation between pre-training data and model performance.

Biomedical pre-training is not sufficient to achieve state-of-the-art performance. As seen in Table 2 3/5 of the top 5 systems for the entailment task and the evidence selection task do not apply any biomedical pre-training strategies. Furthermore, Kanakarajan and Sankarasubbu (2023) @Saama AI Research demonstrates that large generative models are capable of outperforming the majority-class baseline on the entailment task, even in a zero-shot setting. Additionally, Vladika and Matthes (2023) and Wang et al. (2023) record DeBERTa-v3 (He et al., 2021) significantly outperforming comparatively smaller models pre-trained on biomedical data.

7.5 Evidence-based NLI

Many of the discriminative models have a limited input length, often smaller than the CTR premise token length (Alameldin and Williamson, 2023). Therefore extracting a condensed set of evidence facts, prevents the information from being lost by truncation. Even for generative models adapted to receive longer sequences of text, there is still a risk of distractors present in the CTR premise interfering with the inference process, particularly with respect to numerical inference.

Retrieving evidence before inference does not result in better entailment task performance. Systems that execute the evidence selection task, extract relevant evidence from the premise with respect to the statement, and then use the retrieved evidence for the entailment task (Pre), do not demonstrate significantly higher F1 than models which perform the inference over the entire premise (Post), shown in Table 2. As mentioned previously the cost of excluding relevant information is significant, and systems that perform inference over the entire premise circumvent this cost as they effectively have an evidence extraction Recall of 1.0 at the inference step.

Retrospective evidence retrieval induces confirmation bias. 11 participants submit to both tasks, and 6 participants opt to classify the entailment, then retrieve evidence from the premise to support the classification. Conversely, 5 participants first extract relevant evidence and then classify the entailment based on selected evidence. There is no significant difference in the results of these clusters for the entailment task, however for the evidence selection task, systems that first collect evidence average +0.045 F1, and +0.07 Precision compared to those that retrospectively select evidence. These clusters report identical average Recall. Therefore, we hypothesize that retrospective systems exhibit confirmation bias, as selected evidence must be relevant to both the statement and the predicted label. The expected effects of reducing the input size by filtering out irrelevant parts of the premise are not evident in the reported results.

7.6 Limitations

Joint inference systems may generalize poorly without prior knowledge. NLI4CT was constructed using a negative-rewriting strategy (Section 5), this results in one contradictory statement and one entailment statement for each CTR premise. Alissa and Abdullah (2023) @JUST-KM and Zhou et al. (2023) @THiFLY leverage this feature. These systems perform inference over statement pairs with shared premises and assign the entailment label to the statement with the highest confidence, and then assign the contradiction label to the remaining statement, regardless of confidence. Zhou et al. (2023) @THiFLY reports that this process improves entailment task performance by +0.8 F1. The limitations of this approach are that this is heavily reliant on the knowledge that only one statement is entailed, and therefore this approach may generalize poorly where this knowl-

edge is unavailable.

8 Statistical Artifacts

Statistical characteristics such as imbalanced sequence lengths, token distributions, or discriminative conditions that are disproportionately associated with a particular class can superficially inflate model performance (Herlihy and Rudinger, 2021). Alameldin and Williamson (2023) @Clemson NLP observes that systems are able to outperform the random baseline on the entailment task using only the statements, this indicates that systems are able to exclusively rely on the presence of superficial statistical patterns within the collection of statements, without learning the underlying rules of the tasks, reporting an F1 of 0.584. This is significantly below the majority baseline (0.66 F1), and as the entailment task is a binary classification task we conclude that the effects of these artifacts on the submitted results are very minimal.

Alameldin and Williamson (2023) @Clemson NLP identifies minor differences in statement lengths across classes, however, in our analysis we did not find a significant difference. Additionally, we retrieve the 15 most frequently used tokens in the statements for both classes, although we observed some uneven distributions in the training set, these distributions were not present or even inversely correlated in the test set. Therefore, we do not believe either of these characteristics explains the observed results, and claim that NLI4CT is robust to statistical biases.

9 Conclusion

This paper presents the systems and results submitted to the SemEval-2023 Task 7 on the NLI4CT dataset. The tasks are challenging, with the majority of submitted systems failing to significantly outperform the majority class baseline on the entailment task. Potentially due to the requirement for sophisticated numerical reasoning, an elevated frequency of biomedical expressions, or the relatively small training set. We observe significantly better performance on the evidence selection task than on the entailment task, and we find that there is no consistent correlation between task performances. The impact of biomedical pre-training is significantly less profound than expected, far outweighed by the effects of increased model size. There is a direct correlation between model size and task performance, with MLMs achieving the highest results in both tasks. There remains room for improvement on both tasks, potentially by exploiting data augmentation to increase training set size, leveraging the zero-shot capabilities of models such as GPT and T5, or through the direct integration of domain knowledge from ontologies. A further error analysis is necessary to evaluate the impact of biomedical pre-training on MLMs, consistency of performance across sections, generalization ability of models trained on NLI4CT, and comparison of performance on numerical versus biomedical instances.

References

Ahamed Alameldin and Ashton Williamson. 2023. Clemson NLP at SemEval-2023 Task 7: Applying GatorTron to Multi-evidence Clinical NLI. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Kefah Alissa and Malak Abdullah. 2023. JUST-KM at SemEval-2023 Task 7: Multi-evidence Natural Language Inference using Role-based Double Roberta-Large. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Nancy E Avis, Kevin W Smith, Carol L Link, Gabriel N Hortobagyi, and Edgardo Rivera. 2006. Factors associated with participation in breast cancer treatment clinical trials. *J Clin Oncol*, 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.

Robert Bevan, Oisín Turbitt, and Mouhamad Aboshokor. 2023. MDC at SemEval-2023 Task 7: Fine-tuning Transformers for Textual Entailment Prediction and Evidence Retrieval in Clinical Trials. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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*, Toronto, Canada. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *CoRR*, abs/1909.02164.

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.

Sofia I. R. Conceição, Diana F. Sousa, Pedro M. Silvestre, and Francisco M Couto. 2023. lasigeBioTM at SemEval-2023 Task 7: Improving Natural Language Inference Baseline Systems with Domain Ontologies. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Abel Corrêa Dias, Filipe Faria Dias, Higor Moreira, Viviane Moreira, and João Luiz Dihl 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*, Toronto, Canada. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Eric P. Lehman, Benjamin E. Nye, Iain James Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. *ArXiv*, abs/2005.04177.

Chao Feng, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 Task7: Multi-evidence Natural Language Inference for Clinical Trial Data Based a BioBERT Model. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Alexandre Galashov, Jonathan Schwarz, Hyunjik Kim, Marta Garnelo, David Saxton, Pushmeet Kohli, S. M. Ali Eslami, and Yee Whye Teh. 2019. Meta-learning surrogate models for sequential decision making. *CoRR*, abs/1903.11907.

Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. 2021. A deep database of medical abbreviations and acronyms for natural language processing. *Scientific Data*, 8(1):1–9.

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 Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Christine Herlihy and Rachel Rudinger. 2021. Mednli is not immune: Natural language inference artifacts in the clinical domain. *arXiv preprint arXiv:2106.01491*.

Mingtong Huang, Junxiang Ren, Lang Liu, Ruilin Song, and Wenbo Yin. 2023. CPIC at SemEval-2023 Task 7: GPT2-based Model for Multi-evidence Natural Language Inference for Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.

Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):1–14.

Qiao Jin, Jinling Liu, and Xinghua Lu. 2019. Deep contextualized biomedical abbreviation expansion. *arXiv preprint arXiv:1906.03360*.

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*, Toronto, Canada. Association for Computational Linguistics.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

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. *CoRR*, abs/1901.08746.

Rahmad Mahendra, Damiano Spina, and Karin Verspoor. 2023. ITTC at SemEval 2023-task 7: Document Retrieval and Sentence Similarity for Evidence Retrieval in Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Swaroop Mishra and Bhavdeep Singh Sachdeva. 2020. Do we need to create big datasets to learn a task? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 169–173, Online. Association for Computational Linguistics.

Sungrim Moon, Bridget McInnes, and Genevieve B Melton. 2015. Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. *Healthcare informatics research*, 21(1):35–42.

Mariana Neves. 2023. Bf3R at SemEval-2023 Task 7: A Text Similarity Model for Textual Entailment and Evidence Retrieval in Clinical Trials and Animal Studies. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Sheerin Sitara Noor Mohamed and Kavitha Srinivasan. 2023. SSNSheerinKavitha at SemEval-2023 Task 7: Semantic Rule Based Label Prediction Using TF-IDF and BM25 Techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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*, Toronto, Canada. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? *CoRR*, abs/2103.07191.

Ahmad Pesaranghader, Stan Matwin, Marina Sokolova, and Ali Pesaranghader. 2019. deepbiowsd: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association*, 26(5):438–446.

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. *CoRR*, abs/2106.03598.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Saravanan Rajamanickam and Kanagasabai Rajaraman. 2023. I2R at SemEval-2023 Task 7: Explanations-driven Ensemble Approach for Natural Language Inference over Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Penstein Rosé, and Eduard H. Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. *CoRR*, abs/1901.03735.

Alexey Romanov and Chaitanya Shivade. 2018a. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Alexey Romanov and Chaitanya Shivade. 2018b. Lessons from natural language inference in the clinical domain. *CoRR*, abs/1808.06752.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Ian Soboroff. 2021. Overview of trec 2021. In *30th Text REtrieval Conference. Gaithersburg, Maryland*.

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):1–10.

Conner Takehana, Dylan Lim, Emirhan Kurtulus, Ramya Iyer, Ellie Tanimura, Pankhuri Aggarwal, Molly Cantillon, Alfred Yu, Sarosh Khan, Nathan Chi, and Ryan Chi. 2023. Stanford MLab at SemEval 2023 Task 7: Neural Methods for Clinical Trial Report NLI. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, Deborah Ferreira, Julia Rozanova, and André Freitas. 2022. Diff-explainer: Differentiable convex optimization for explainable multi-hop inference. *Transactions of the Association for Computational Linguistics*, 10:1103–1119.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. Unification-based reconstruction of multi-hop explanations for science questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 200–211.

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*, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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*, Toronto, Canada. Association for Computational Linguistics.

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*, Toronto, Canada. Association for Computational Linguistics.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023. KnowComp at SemEval-2023 Task 7: Finetuning Pre-trained Language Models for Clinical Trial Entailment Identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176.

Xiaofeng Zhao, Min Zhang, Miaomiao Ma, Chang Su, Minghan Wang, Xiaosong Qiao, Jiaxin Guo, Yinglu Li, Wenbing Ma, Shimin Tao, and Hao Yang. 2023. HW-TSC at SemEval-2023 Task 7: Exploring the Natural Language Inference Capabilities of ChatGPT and PretrainedLM for Clinical Trial. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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*, Toronto, Canada. Association for Computational Linguistics.