# SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials

**Maël Jullien[1], Marco Valentino[3], André Freitas[1,2,3]**
[1]Department of Computer Science, University of Manchester, UK
[2] National Biomarker Centre, CRUK-MI, University of Manchester, UK
[3]Idiap Research Institute, Switzerland
[1]{firstname.surname}@manchester.ac.uk
[3]{firstname.surname}@idiap.ch

## Abstract

Large Language Models (LLMs) are at the forefront of NLP achievements but fall short in dealing with shortcut learning, factual inconsistency, and vulnerability to adversarial inputs. These shortcomings are especially critical in medical contexts, where they can misrepresent actual model capabilities. Addressing this, we present SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. Our contributions include the refined NLI4CT-P dataset (i.e. Natural Language Inference for Clinical Trials - Perturbed), designed to challenge LLMs with interventional and causal reasoning tasks, along with a comprehensive evaluation of methods and results for participant submissions. A total of 106 participants registered for the task contributing to over 1200 individual submissions and 25 system overview papers. This initiative aims to advance the robustness and applicability of NLI models in healthcare, ensuring safer and more dependable AI assistance in clinical decision-making. We anticipate that the dataset, models, and outcomes of this task can support future research in the field of biomedical NLI. The dataset[1], competition leaderboard[2], and website[3] are publicly available.

## 1 Introduction

Large Language Models (LLMs) excel in numerous Natural Language Processing (NLP) tasks, as evidenced by their state-of-the-art achievements (Brown et al., 2020; Chowdhery et al., 2022). Despite these advancements, LLMs are prone to several critical vulnerabilities. These include a tendency towards shortcut learning, which may compromise their learning process and accuracy (Geirhos et al., 2020; Poliak et al., 2018; Tsuchiya, 2018). Additionally, they exhibit factual inconsistencies (Elazar et al., 2021) and are sensitive to changes in word distributions (Miller et al., 2020; Lee et al., 2020), data transformations (Xing et al., 2020; Stolfo et al., 2022; Meadows et al., 2023; Rozanova et al., 2023), and adversarial attacks (Li et al., 2020). These issues are particularly concerning as they may lead to an overestimation of LLMs' capabilities in practical applications, a risk that is notably significant in fields requiring high reliability, such as healthcare (Patel et al., 2008; Recht et al., 2019).

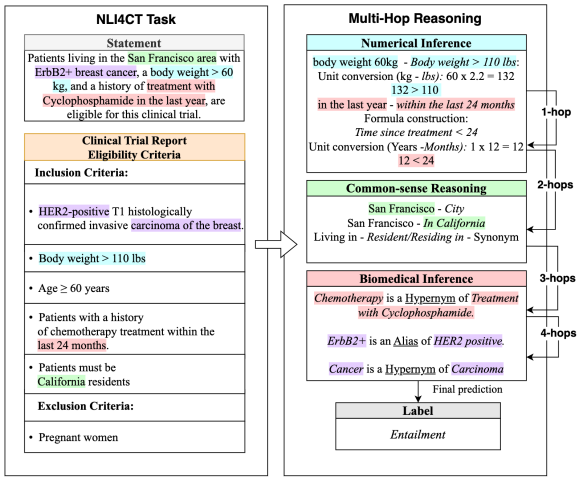


Figure 1: The goal of NLI4CT is to predict the relationship of entailment between a **Statement** and a **CTR** premise (Jullien et al., 2023a). In this task, we introduce a set of perturbations (NLI4CT-P) applied to the statements to test the semantic consistency and faithfulness of NLI models.

Clinical trials play a pivotal role in evaluating the efficacy and safety of novel treatments, thereby significantly contributing to the progress of experimental medicine (Avis et al., 2006). Clinical Trial Reports (CTRs) document the methodologies and outcomes of these trials, serving as a foundation for healthcare professionals to devise and administer experimental therapies. However, the sheer volume

[1]https://github.com/ai-systems/nli4ct
[2]https://codalab.lisn.upsaclay.fr/competitions/16190
[3]https://sites.google.com/view/nli4ct/



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962

of CTRs, exceeding 400,000 and continually growing (Bastian et al., 2010), renders it impractical for a manual comprehensive analysis of all pertinent literature in treatment planning (DeYoung et al., 2020). In this context, Natural Language Inference (NLI) (Bowman et al., 2015) emerges as a viable solution, facilitating the large-scale interpretation and synthesis of medical evidence. This approach effectively bridges the latest research findings with clinical practice, thereby supporting the delivery of personalized care (Sutton et al., 2020).

Previously, we created the Multi-Evidence Natural Language Inference for Clinical Trial Reports (NLI4CT) dataset, detailed in Jullien et al. (2023a). This dataset, enriched with Clinical Trial Reports (CTRs) and expert-annotated statements for entailment and contradiction, exemplified in Figure 1, served as the foundation for organizing "SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data".

While the preceding version of NLI4CT spurred the creation of models based on Large Language Models (LLMs) (Zhou et al., 2023; Kanakarajan and Sankarasubbu, 2023; Vladika and Matthes, 2023) that demonstrated commendable performance (i.e., F1 score $\approx 85\%$), deploying LLMs in sensitive areas like real-world clinical trials mandates additional scrutiny. This necessitates the invention of new evaluation frameworks that allow thorough behavioural and causal analysis (Wang et al., 2021).

In pursuit of these goals, we present the latest iteration of our dataset, NLI4CT-P, an extension of the original NLI4CT with data perturbations. Moreover, we provide a comprehensive analysis of the systems that participated in "SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials" a task conducted using the NLI4CT-P dataset. This initiative aims to improve our understanding of LLMs behaviour and advance evaluation methodologies for clinical Natural Language Inference (NLI).

The task is structured around the systematic application of controlled interventions, each designed to investigate specific semantic and numerical inference challenges typical of clinical NLI (see Table 1). The interventions enable a comprehensive evaluation of LLMs' reasoning capabilities within a clinical framework, emphasizing robustness, consistency, and faithfulness.

Our efforts aim to significantly contribute to the crafting of more dependable and insightful evaluation standards and metrics for NLI systems, ensuring their reliability and efficacy in healthcare applications.

**Original Statement:** The primary trial intervention protocol lasts a total of 14 days.
**Label:** Entailment

| Perturbed Statement | Intervention | Type |
|---|---|---|
| The primary clinical trial's intervention treatment plan has a duration of 14 days. | Paraphrase | Preserving |
| The primary clinical trial intervention protocol spans an entire year | Contradiction rephrasing | Altering |
| Lacks energy refers to whether an individual has/had a lack of energy. The primary trial intervention protocol lasts a total of 14 days | Text appended | Preserving |
| The primary trial intervention protocol lasts 2 weeks | Numerical paraphrase | Preserving |
| The primary trial intervention protocol lasts a total of 3 hours | Numerical contradiction | Altering |

Table 1: Example of perturbations applied to the statements with the type of intervention and its semantic effect (i.e., preserving vs. altering).

This second iteration is intended to ground NLI4CT in interventional and causal analyses of NLI models (YU et al., 2022). By enriching the original NLI4CT dataset with a novel contrast set derived from targeted interventions to statements in the NLI4CT test and development sets, we establish a direct causal link between these interventions and the anticipated labels. This enhancement introduces two innovative metrics, Consistency and Faithfulness. These metrics allow us to explore specific research objectives with a causal perspective:

- **Consistency:** To examine whether NLI models maintain uniformity in processing semantically equivalent phenomena crucial for inference within clinical NLI contexts.
- **Faithfulness:** To assess the capacity of NLI models to capture and interpret the underlying semantic features required for reasoning over clinical trials, and to change their predictions according to relevant changes of such features.

This paper introduces SemEval-2024 Task 2 – Safe Biomedical Natural Language Inference for Clinical Trials – (NLI4CT-P) presenting a detailed analysis of the performance of the participating systems. We report the following conclusions:

**Challenges in Clinical NLI:** Despite improvements achieved via the application of Large Language Models (LLMs), Clinical NLI remains a significant challenge. With the highest F1 score achieved in this task being 0.8 (Liu and Thoma, 2024; Guimarães et al., 2024) (FZI-WIM, Lisbon Computational Linguists), leveraging Mixtral-8x7B-Instruct models. This emphasises the necessity for the development of more robust and reliable systems capable of dealing with the challenges of real-world clinical application.

**Importance of Faithfulness and Consistency Evaluation:** The incorporation of Faithfulness and Consistency metrics into our evaluation framework underscores the unpredictability of current systems and the limitations inherent in relying solely on F1 score for comprehensive analysis.

**Superiority of Generative Models:** Generative models have been shown to outperform discriminative models in terms of F1 score (+0.025), Faithfulness (+0.15), and Consistency (+0.037).

**Value of Additional Data:** Leveraging additional training data in the form of instruction tuning or medical NLI datasets has been shown to produce significant performance gains. When augmented with extra data, systems exhibit notable enhancements, recording improvements of +0.056 in F1 score, +0.132 in Faithfulness, and +0.062 in Consistency relative to their counterparts.

**Impact of Prompting Strategies:** The study highlights that the choice of prompting strategy plays a crucial role in influencing model performance. Specifically, zero-shot prompting has been shown to provide notable enhancements, with an average increase of +0.025 in F1 score, and marginal gains of +0.001 in both Faithfulness and Consistency, compared to the outcomes achieved with few-shot prompting techniques.

**Efficacy of Mid-Sized Architectures:** Mid-sized architectures, possessing 7B to 70B parameters, offer a cost-effective alternative capable of matching or surpassing larger models in key performance metrics like F1, Faithfulness, and Consistency. Compared to models exceeding 70B parameters, these mid-sized models report a slight improvement of +0.01 in F1 score, albeit with minor reductions of -0.03 in Faithfulness and -0.01 in Consistency. Against models below 7B parameters, however, they show notable enhancements, achieving +0.10 in F1 score, +0.40 in Faithfulness, and +0.19 in Consistency.

## 2 Task Description

SemEval-2024 Task 2 is a textual entailment task, each instance in NLI4CT contains a CTR premise and a related statement. These premises range from 5 to 500 tokens in length and provide details about a trial's results, eligibility criteria, interventions, or adverse events. Corresponding statements are concise sentences, containing 10 to 35 tokens, that make some claim about the premise information (refer to Table 1 for examples). The task is to classify the inference relation between a CTR premise, and a statement as either entailment or contradiction, exemplified in Figure 1 The dataset features two distinct types of instances: single instances, where a statement discusses a single CTR, and comparison instances, which involve statements that compare and contrast two CTRs.

## 3 Dataset

The premises used in the NLI4CT dataset (Jullien et al., 2023a) are derived from 1,000 publicly accessible, English-language breast cancer Clinical Trial Reports (CTRs) published on ClinicalTrials.gov a resource managed by the U.S. National Library of Medicine. This dataset complies with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. The original NLI4CT collection includes 2,400 expert-annotated statements, premises and associated labels. These are distributed across training, testing, and development sets in a 70/20/10 ratio.

We have advanced the methodology of the previous NLI4CT dataset by incorporating interventions to create a contrast set, enabling a systematic behavioural and causal analysis of models evaluated in the competition. This enhanced version is referred to as NLI4CT-P (Perturbed). The construction of the contrast set involves four semi-automated, controlled interventions applied to the statements from the NLI4CT test and development set. It's important to note that the specifics of these interventions were kept undisclosed until the completion of the competition's testing phase on January 31st 2024.

### 3.1 Interventions

We delineate and implement the four interventions in the following manner:

**Paraphrasing and Contradiction Rephrasing** Clinical texts frequently contain acronyms and aliases, which can hinder the performance of clinical NLI models (Grossman Liu et al., 2021; Jimeno-Yepes et al., 2011; Pesaranghader et al., 2019; Jin et al., 2019). Moreover, these models can fall prey to shortcut learning, where they make inferences based on syntactic patterns rather than semantic understanding (Geirhos et al., 2020). To evaluate this phenomenon, original statements were rephrased using different vocabulary and syntax. Paraphrasing was employed to retain the original meaning and label (row 1 Table 1), while contradiction rephrasing created new statements that directly contradict the original statement, and are therefore always labelled as contradictions (row 2 Table 1).

**Numerical Paraphrasing and Contradiction** Large Language Models (LLMs) have shown limitations in consistent numerical and quantitative reasoning (Patel et al., 2021; Ravichander et al., 2019; Galashov et al., 2019), an essential aspect for tasks like NLI4CT that demand such inferences. To evaluate the models' capabilities in this area, operands and numerical units within the hypotheses were altered (rows 4 and 5 Table 1). This modification either preserved or inverted the initial entailment label.

**Appending Text** LLMs are often challenged by complex reasoning when dealing with extended premise-hypothesis pairs (Liu et al., 2021). We test this in a clinical setting by appending biomedical definitions from the NCI Thesaurus to the original statements (row 3 Table 1). The added definitions, ranging from 15 to 20 tokens in length, almost double the average statement token length. Despite the definitions not being independently verifiable against the premises, these definitions are regarded as 'ground truth', they are universally true and remain neutral in relation to the premises. Since they neither assert nor verify any premise-specific information, within the scope of our task, appending such neutral text is categorized as a 'preserving' intervention.

These interventions, other than the text appending, were performed by prompting ChatGPT 3.5 and Whisper APIs (Brockman et al., 2023) with human-in-the-loop correction to address any errors (Gilardi et al., 2023). Each statement in the test and development sets underwent each type of intervention process three times. This did not extend to the training set, as the aim was to prevent models from learning the patterns of intervention. Although attempts were made to apply numerical paraphrasing and contradiction interventions, they were not always feasible. This was due to the absence of numerical data or units in the original statements, and when the quality of the perturbed statements was deemed substandard, they were excluded during the manual review phase. consequently, this resulted in a markedly reduced count of numerically perturbed statements within the dataset. The prompts used to perform the interventions are available in the appendix.

## 4 Evaluation

SemEval-2024 Task 2 is devised as a binary classification challenge, with the Macro F1-score being utilized to gauge the foundational performance of the participating systems. This evaluation is conducted on the original NLI4CT test set, serving as a control metric, rather than on the NLI4CT-P test set, which contains exclusively perturbed statements. Although the Macro F1 score is instrumental in measuring overall model performance by highlighting precision and recall across various classes, it inherently lacks the capability to fully capture the sophisticated understanding and reasoning skills essential for effective Natural Language Inference (NLI). Specifically, the F1 score does not assess a model's capacity to adjust to subtle semantic shifts or evaluate the resilience of its predictions when faced with interventions that either modify or maintain the semantic integrity of statements. This gap highlights the necessity for more advanced metrics capable of offering deeper insights into a model's interpretative and reasoning proficiency. In response to this need and inspired by recent advancements in causal analysis within the NLP domain (Stolfo et al., 2022), we introduce two novel evaluation metrics aimed at examining the causal effects of interventions on model performance.

**Faithfulness** gauges the degree to which a system's predictions are both accurate and grounded in the correct rationale. Intuitively, this is estimated by measuring the ability of a model to correctly adjust its predictions when exposed to interventions that modify the meaning (semantic altering) of the statement. Specifically, for a set of $N$ statements $x_i$ in the contrast set ($C$), alongside their corresponding original statements $y_i$ and the model predictions denoted as $f()$, faithfulness is quantified using the

| Set | Original | Appended definition | Paraphrase | Contradiction rephrasing | Numerical paraphrase | Numerical Contradiction | Total |
|---|---|---|---|---|---|---|---|
| **Dev** | 200 | 600 | 600 | 600 | 64 | 78 | 1942 |
| **Test** | 500 | 1500 | 1500 | 1500 | 224 | 276 | 5000 |

Table 2: Distribution of statement counts across the sets of NLI4CT-P

formula presented in Equation 1.

$$Faithfulness = \frac{1}{N}\sum_{1}^{N}|f(y_i) - f(x_i)|$$
$$x_i \in C : \text{Label}(x_i) \neq \text{Label}(y_i), \text{ and } f(y_i) = \text{Label}(y_i) \quad (1)$$

**Consistency** assesses a system's capability to generate identical outcomes for semantically equivalent inputs. This measure evaluates whether a system can uniformly predict the same label for both the original and contrast statements under interventions that do not alter the semantic content (semantic preserving) of the statements. The key aspect here is the uniformity in representing semantic concepts across different statements, irrespective of the correctness of the final prediction. For $N$ statements $x_i$ in the contrast set ($C$), alongside their original counterparts $y_i$, and model predictions $f()$, consistency is determined as follows:

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

The Macro F1 score provides a foundational benchmark for basic model performance, serving as a control metric, the core objective of Task 2 is towards enhancing model quality and dependability through systematic causal analysis. The pursuit here is not only for high performance in a traditional sense but for models that demonstrate a more reliable and robust application of natural language, reflecting a more nuanced approach to evaluating system capabilities, and allowing for developing safer, ethical, and trustworthy clinical systems.

## 5 Results and Discussion

106 participants registered to the SemEval-2024 Task 2 competition contributing over 1200 individual submissions and 25 system overview papers, presented in Table 3. Please note that our analysis focuses exclusively on systems that are detailed in system overview papers and for which official leaderboard results have been provided. Generally, participants tend to submit the highest-scoring results to the leaderboard, regardless of whether the system achieving these results represents the primary contribution of their paper. This approach ensures that our report reflects the peak performance levels achieved, albeit potentially overlooking the main systems of interest described in the papers.

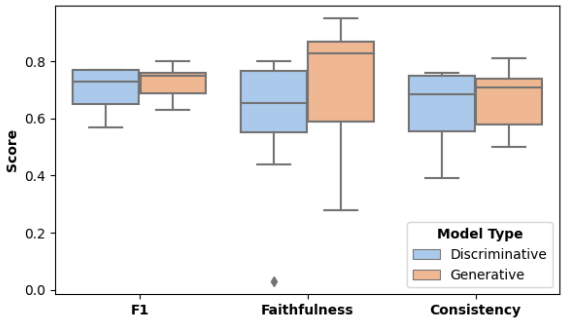


Figure 2: Comparative Analysis of F1, Consistency, and Faithfulness Across Model Types

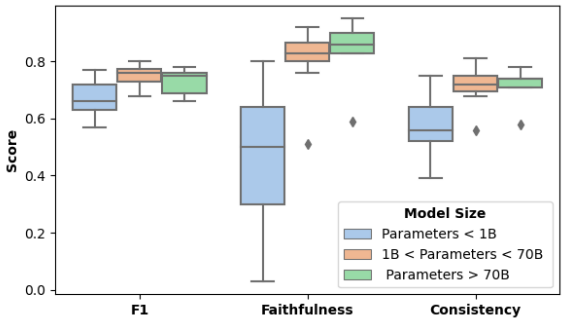


Figure 3: Comparative Analysis of F1, Consistency, and Faithfulness Across Model Parameter Numbers

### 5.1 Architectures

In the SemEval-2024 Task 2 submissions, a diverse range of 12 different architectures was employed, as detailed in Table 4. The predominant choice among participants was Mistral-based architectures, accounting for 7 out of 25 submissions, closely followed by DeBERTa with 5 out of 25. The majority of submissions utilised generative

| Work | F1 | F | C | Average Score | Architecture | Inference Strategies | Fine-Tuning | Dataset Augmentation |
|---|---|---|---|---|---|---|---|---|
| FZI-WIM (Liu and Thoma, 2024) | **0.8** | 0.9 | 0.73 | 0.81 | Mixtral-8x7B-Instruct | CoT | Yes | GPT-4, bart-large-mnli Instruction Dataset |
| Lisbon Computational Linguists (Guimarães et al., 2024) | **0.8** | 0.83 | 0.72 | 0.78 | Mistral-7B-Instruct-v0.2 | Zero-shot | Yes | Mistral-7B-Instruct-v0.2 dataset expansion |
| NYCU-NLP (Lee et al., 2024) | 0.78 | 0.92 | **0.81** | **0.84** | SOLAR (10.7B) | Zero-shot | Yes | OpenChat v3.5, Intervention Reduction |
| Edinburgh Clinical NLP (Gema et al., 2024) | 0.78 | **0.95** | 0.78 | **0.84** | GPT-4 | Zero-shot | No | - |
| YNU-HPCC (Zhang et al., 2024) | 0.77 | 0.67 | 0.73 | 0.72 | DeBERTa-v3-large | Discriminative | Yes | MultiNLI, FeverNLI, ANLI, LingNLI, WANLI, Back Translation |
| BD-NLP (Nath and Samin, 2024) | 0.77 | 0.79 | 0.76 | 0.77 | DeBERTa-lg | Discriminative | Yes | - |
| CaresAI (Abdel-Salam et al., 2024) | 0.77 | 0.76 | 0.75 | 0.76 | Ensemble of DeBERTas | Discriminative | Yes | - |
| TüDuo (Smilga and Alabiad, 2024) | 0.76 | 0.84 | 0.75 | 0.78 | Flan-T5 XL | Few-shot | Yes | GPT-3.5-Turbo Instruction Dataset |
| RGAT (Chakraborty, 2024) | 0.76 | 0.86 | 0.74 | 0.79 | GPT-4 | Zero-shot | No | - |
| DFKI-NLP (Verma and Raithel, 2024) | 0.75 | 0.81 | 0.68 | 0.75 | Mistral 7B | Zero-shot | Yes | Meta-Inventory dataset expansion, MedNLI |
| D-NLP (ALTINOK, 2024) | 0.75 | 0.83 | 0.74 | 0.77 | Gemini Pro | Zero-shot | No | - |
| LMU-BioNLP (Sun et al., 2024) | 0.75 | 0.86 | 0.69 | 0.77 | Mistral-7b | Zero-shot | Yes | GPT-3.5, GPT4 dataset expansion, and instruction tuning dataset |
| DKE-Research (Wang et al., 2024) | 0.74 | 0.8 | 0.75 | 0.76 | DeBERTa-l | Discriminative | Yes | GPT-3.5, TF-IDF dataset expansion |
| Puer (Dao et al., 2024) | 0.72 | 0.59 | 0.64 | 0.65 | Biolinkbert-large | Discriminative | Yes | - |
| UniBuc (Micluța-Câmpeanu et al., 2024) | 0.71 | 0.83 | 0.72 | 0.75 | SOLAR 10B | few-shot | No | - |
| iML (Akkasi et al., 2024) | 0.7 | 0.28 | 0.52 | 0.50 | SciFive | Zero-shot | Yes | - |
| CRCL (Brutti-Mairesse, 2024) | 0.7 | 0.87 | 0.7 | 0.76 | Mixtral-8x7B | CoT, OPRO optimization | No | - |
| IITK (Mandal and Modi, 2024) | 0.69 | 0.9 | 0.71 | 0.77 | Gemini Pro | Zero-shot, ToT and CoT | No | - |
| 0x.Yuan (Lu and Kao, 2024) | 0.68 | 0.51 | 0.56 | 0.58 | Mixtral-8x7B | multi-agent debating framework | No | - |
| Saama Technologies (Kim et al., 2024) | 0.66 | 0.59 | 0.58 | 0.61 | Gemini Pro, mistral-7B-instruct-v0.2 | CoT, Few-Shot | Yes | - |
| TLDR (Das et al., 2024) | 0.66 | 0.5 | 0.58 | 0.58 | SciFive-base, DeBERTa-v3-base | Zero-shot | No | - |
| Concordia University (Marks et al., 2024) | 0.66 | 0.03 | 0.39 | 0.36 | BART | Discriminative | Yes | - |
| T5-Medical (Siino, 2024) | 0.63 | 0.3 | 0.5 | 0.48 | T5-large-medical | Zero-Shot | No | - |
| USMBA-NLP (Fahfouh et al., 2024) | 0.62 | 0.44 | 0.54 | 0.53 | BERT base | Discriminative | Yes | - |
| SEME (Aguiar et al., 2024) | 0.57 | 0.64 | 0.56 | 0.59 | NLI-RoBERTa ensemble | Discriminative | Yes | - |

Table 3: SemEval-2024 Task 2 Results, sorted by F1 (on the unperturbed subset of the test set), with Faithfulness (F), and Consistency (C)

models, with 17 out of the total, compared to 8 leveraging discriminative models. The F1 score suggests that GPT-4's performance is on par with considerably smaller models such as DeBERTa. However, a deeper evaluation using our novel metrics, especially Faithfulness, reveals a significant disparity, indicating that smaller models might be overfitting. This observation underscores the importance of employing these complementary metrics for a more comprehensive comparison of model capabilities. Despite the prevailing notion that larger models inherently perform better, this trend appears to be less pronounced than observed in this task's previous iteration (Jullien et al., 2023b), as illustrated in Figure 3. Notably, there seems to be a point of diminishing returns for model sizes between 7B and 70B, within the generative model category, shown in Figure 3. On average, models with sizes ranging from 7B to 70B parameters achieve +0.01 in F1 score but show decreases of -0.03 in Faithfulness and -0.01 in Consistency relative to models with more than 70B parameters. When compared to models with fewer than 7B parameters, these mid-sized models exhibit substantial improvements of +0.10 in F1 score, +0.40 in Faithfulness, and +0.19 in Consistency.

Table 4: Participant architectures by popularity, with average F1, Faithfulness (F) and Consistency (C)

| Model | F1 | F | C | Count |
|---|---|---|---|---|
| DeBERTa | 0.76 | 0.76 | 0.75 | 5 |
| Mistral 7B | 0.75 | 0.84 | 0.69 | 4 |
| Mixtral 8x7B | 0.73 | 0.76 | 0.66 | 3 |
| T5 | 0.66 | 0.36 | 0.53 | 3 |
| Gemini Pro | 0.70 | 0.77 | 0.68 | 3 |
| GPT-4 | **0.77** | **0.91** | 0.76 | 2 |
| SOLAR 10B | 0.75 | 0.88 | **0.77** | 2 |
| BERT base | 0.62 | 0.44 | 0.54 | 1 |
| Biolinkbert | 0.72 | 0.59 | 0.64 | 1 |
| BART | 0.66 | 0.03 | 0.39 | 1 |
| RoBERTa | 0.57 | 0.64 | 0.56 | 1 |
| Flan-T5 XL | 0.76 | 0.84 | 0.75 | 1 |

Additionally, on average, generative models outperform discriminative ones across the board—with improvements observed in F1 scores (+0.025), Faithfulness (+0.15), and Consistency (+0.037), as depicted in Figure 2. Intriguingly, when comparing specific architectures, there is minimal correlation between model types and Faithfulness, Consistency, and F1, even though the top two performing systems in terms of F1 score are based on the Mixtral-8x7B-Instruct model (see Table 3).

### 5.2 Base F1 Performance

As previously mentioned the focus of this task extends beyond base performance. Nevertheless, it's noteworthy that the highest F1 score achieved in this iteration was 0.8 (Liu and Thoma, 2024; Guimarães et al., 2024) (FZI-WIM, Lisbon Computational Linguists) by two systems (Table 3). A figure that notably falls short of the previous iteration's top score of 0.856 (Zhou et al., 2023; Jullien et al., 2023b). This observed decline underscores a significant gap between the current capabilities of NLI systems and the performance required for practical application within clinical environments.

### 5.3 Faithfulness and Consistency

The overall average Faithfulness recorded at 0.719 significantly outperforms the average Consistency, which stands at 0.67. This disparity grows more pronounced within the subset of models within the top 10 F1 scores, where Average Faithfulness escalates to 0.835 and Average Consistency to 0.751.

Furthermore, a robust overall Spearman's correlation was identified between Consistency and F1 scores (0.8) and between Faithfulness and F1 scores (0.62). Intriguingly, this correlation inverts within the top 10 systems, where Spearman's Correlation between Consistency and F1 drops to -0.12, and between Faithfulness and F1 rises slightly to 0.319. Notably, the models with the highest Faithfulness (0.95) (Gema et al., 2024)(Edinburgh Clinical NLP) and Consistency (0.81) (Lee et al., 2024)(NYCU-NLP) scores achieve an average score of 0.84, surpassing systems ranked above them (with average scores of 0.81 and 0.78) yet both reporting a lower F1 score by -0.02. also Mandal and Modi (2024)(IITK) achieves a very high faithfulness of 0.9, while only managing an F1 of 0.69. These patterns underscore the limitation of F1 scores as sole indicators of model performance at the apex levels, accentuating the importance of considering Faithfulness and Consistency metrics in conjunction with F1.

The inversion of correlations among the top 10 models suggests a nuanced landscape of performance evaluation. While Consistency contributes broadly to high F1 scores, the top 10 models distinctly leverage Faithfulness, indicating that, at peak performance levels, perhaps accurate predictions rooted in correct premises are paramount over consistent responses to similar cases.

This phenomenon might also signify a ceiling effect for Consistency, suggesting that beyond a certain point, efforts to improve consistency do not translate into proportional performance gains. Such a scenario could inadvertently overshadow other critical model attributes like adaptability and nuanced comprehension, aspects more closely associated with Faithfulness. Alternatively, this situation could imply that models specifically optimized for F1 scores might inadvertently neglect Consistency, and to some degree, Faithfulness, as evidenced by the observed decline in their correlation with peak F1 scores.

Our analysis further elucidates the relationship between Consistency and Faithfulness in submitted systems, revealing an Overall Spearman Correlation of 0.708. This correlation slightly diminishes among the top 10 F1 scoring models to 0.39. While this represents a weaker correlation within the subset of the top 10 models, it importantly suggests the absence of a strict trade-off between Consistency and Faithfulness. Such a finding challenges the notion that improvements in one metric necessarily come at the expense of the other.

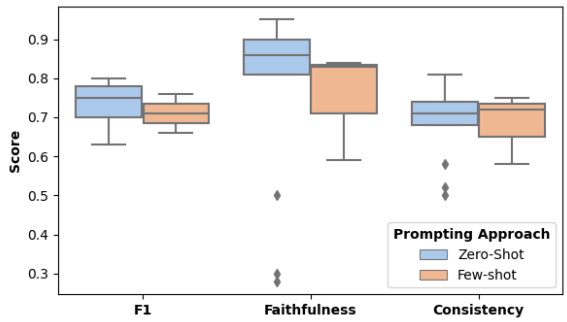


Figure 4: Comparative Analysis of F1, Consistency, and Faithfulness Across Prompting strategies

Among the participants, 4 out of 25 achieved a Faithfulness score of 0.9 or higher (Mandal and Modi, 2024; Liu and Thoma, 2024; Lee et al., 2024; Gema et al., 2024)(IITK, FZI-WIM, NYCU-NLP, Edinburgh Clinical NLP). Remarkably, only 1 out of 25 participants attained a Consistency score of 0.8 or higher (Lee et al., 2024)(NYCU-NLP). These results suggest a continued need for refining these models to achieve higher degrees of Faithfulness and Consistency if they are to be applied in real-world clinical environments.

### 5.4 Prompting Strategies

A variety of prompting strategies were used in the submitted systems. It is essential to acknowledge that variations in prompts can lead to significant differences in outcomes, even when employing the same architecture. For instance, within the Gemini Pro systems, a comparison between submissions by ALTINOK (2024)(D-NLP) and Kim et al. (2024)(Saama Technologies) from Saama Technologies reveals substantial disparities in performance metrics: F1 scores, Faithfulness, and Consistency differ by 0.09, 0.24, and 0.16, respectively. Similar patterns of variation were observed among submissions utilizing Mistral-based and T5-based approaches, underscoring the impact of prompting nuances.

Among the generative model submissions, 13 out of 16 employed a zero-shot approach, while the remaining three opted for few-shot prompting. Zero-shot prompting involves generating responses without any example-based guidance, relying solely on the model's pre-existing knowledge and the task description. Few-shot prompting, on the other hand, provides the model with one or more examples to guide its responses, traditionally anticipated to yield superior results.

Contrary to initial expectations, zero-shot prompting has shown a significant advantage, especially in achieving higher F1 scores and improving Faithfulness. Notably, four out of the top five models with the highest F1 scores utilized zero-shot techniques, as depicted in Figure 4. On average, zero-shot prompting yielded improvements of +0.025 in F1 score, +0.001 in Faithfulness, and +0.001 in Consistency, when compared to few-shot prompting methods.

Direct prompting is a straightforward method of querying a Language Model (LM). It involves posing a question to the model in a direct manner, without providing additional context or requesting intermediate steps. For example *"Given the CTR:* {Premise} *does the statement:* {Statement} *follow?"*

On the other hand, Chain of Thought (CoT) prompting represents a more elaborate technique designed to prompt the model to "show its work" by articulating the intermediate steps or reasoning that leads to its conclusion (Wei et al., 2022). This approach enables the model to break down the problem into smaller, more manageable parts, thereby facilitating more accurate or explainable predictions. For instance, the prompt could be structured as follows: *"Given the CTR:* {Premise} *and the statement:* {Statement}*, provide a step-by-step reasoning process to determine if the statement logically follows from the report."* Such a modification in the prompting strategy has been shown to produce significant differences in the model's outputs (Wei et al., 2022).

While direct prompting has been the predominant strategy among generative approaches, several teams have experimented with more nuanced strategies. Specifically, FZI-WIM (Liu and Thoma, 2024), IITK (Mandal and Modi, 2024), and Saama Technologies (Kim et al., 2024) have employed Chain of Thought prompting. Furthermore, IITK (Mandal and Modi, 2024) has also explored Tree of Thought (ToT) prompting. ToT prompting is an advanced technique aimed at improving the performance and interpretability of LMs, particularly in complex problem-solving tasks (Yao et al., 2023). It goes beyond the CoT approach by not merely listing reasoning steps linearly but by organizing these steps into a tree structure that represents different branches of reasoning or possible solutions. IITK (Mandal and Modi, 2024) applies this technique with the prompt *Imagine three different clinical experts are answering the question given below. All*

*experts will write down first step of their thinking, then share it with the group. Then all experts will go on to the next step of their thinking. If any expert realises they're wrong at any point then they leave. They will continue till a definite conclusion is reached.*. However, the ability to draw definitive conclusions about the relative efficacy of these prompting strategies is constrained given the considerable performance variability associated with each approach and the application of these strategies across diverse models, complicating efforts to ascertain the sources of performance gains or losses.

Two particularly intriguing prompting strategies emerged from the submissions. (Brutti-Mairesse, 2024)(CRCL) utilized an OPRO (Optimal Prompting for Response Optimization) technique (Yang et al., 2023), which leverages the model's ability to generate effective prompts from a small set of exemplars and prior instructions. This technique essentially tasks the model with creating its own instructions to tackle given problems. Additionally, (Lu and Kao, 2024) introduced a multi-agent debating framework, incorporating several custom agents with diverse expertise, including Biostatistics and Medical Linguistics, to enrich the model's output.

In summary, the submissions reveal a broad spectrum of prompting strategies, from zero-shot to more complex approaches like Tree of Thought and multi-agent frameworks. These strategies significantly influence model performance, underscoring the importance of prompt design in the development and evaluation of NLI systems. As the field progresses, further research is warranted to elucidate the optimal prompting strategies for enhancing model accuracy, reliability, and interpretability across various applications, in a controlled manner.

### 5.5 Fine-tuning strategies

Within the context of SemEval-2024 Task 2, a diverse array of fine-tuning strategies was employed across the 25 participating systems, revealing significant insights into their impact on model performance. Notably, 9 out of 25 systems, all of which were generative, did not undergo any form of fine-tuning. In contrast, 8 out of 25 systems were fine-tuned specifically on the NLI4CT-P training set, while the remaining 6 systems benefited from fine-tuning on additional datasets.

Interestingly, systems fine-tuned on the NLI4CT-P training set exhibited the lowest average performance across all three evaluated metrics, as detailed in Figure 5. Conversely, systems that underwent fine-tuning on external datasets demonstrated superior performance on all metrics, indicating a significant advantage of incorporating diverse training data.

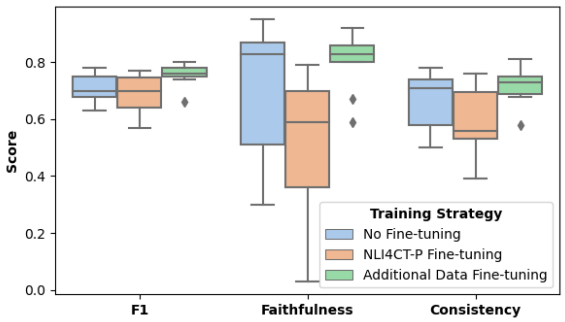


Figure 5: Comparative Analysis of F1, Consistency, and Faithfulness Across Training Strategies

The range of additional datasets leveraged for fine-tuning included various medical NLI datasets, such as MultiNLI, FeverNLI, ANLI, LingNLI, and WANLI, utilized by Zhang et al. (2024)(YNU-HPCC), and MedNLI by Verma and Raithel (2024)(DFKI-NLP). Moreover, some teams, including Sun et al. (2024)(LMU-BioNLP), Wang et al. (2024)(DKE-Research), Guimarães et al. (2024)(Lisbon Computational Lin- guists), Smilga and Alabiad (2024)(TüDuo), and Zhang et al. (2024)(YNU-HPCC), innovatively generated their data by applying interventions similar to those used in our task, thereby enriching their training material. Systems enhanced with additional data demonstrate significant improvements, achieving gains of +0.056 in F1 score, +0.132 in Faithfulness, and +0.062 in Consistency. These results suggest a substantial benefit from such tuning, particularly in terms of Faithfulness. This indicates that incorporating perturbed data into the training process not only enhances the model's inference ability but also significantly improves its reliability and adherence to the truthfulness of the clinical data it processes.

Instruction tuning emerged as a prevalent strategy, with datasets specifically crafted for this purpose by teams such as Liu and Thoma (2024)(FZI-WIM), Guimarães et al. (2024)(Lisbon Computational Lin- guists), Smilga and Alabiad (2024)(TüDuo), LUM-BIO, Wang et al. (2024)(DKE-Research), and (Lee et al., 2024)(NYCU-NLP). Notably, 3 out of the top 5 systems, as per F1 scores, employed instruction tuning, underscoring its effec-

tiveness in enhancing model performance, although notably producing minimal gains in consistency.

## 6 Related Work

The landscape of expert-annotated resources for clinical NLP is rich, with notable examples such as the TREC 2021 Clinical Track (Soboroff, 2021), which focuses on information retrieval from CTR data, highlighting eligibility criteria. Evidence Inference 2.0 (DeYoung et al., 2020) introduces a QA task alongside span selection based on CTR results, while the MEDNLI dataset (Romanov and Shivade, 2018) offers an entailment task using patient medical history notes. These datasets primarily aim to evaluate biomedical language understanding and reasoning. Despite neural architectures leading in biomedical NLI performance (Gu et al., 2021; DeYoung et al., 2020), challenges remain in quantitative reasoning and numerical operations within NLI (Ravichander et al., 2019; Galashov et al., 2019). Prior works experiment with biomedical pre-training strategies (Lee et al., 2020; Shin et al., 2020; Gu et al., 2021), and while ExaCT (Kiritchenko et al., 2010) automates information extraction from clinical trials, the integration of biomedical and numerical NLI effectively remains unaddressed. None of the aforementioned resources provide avenues for meaningful causal analysis, a gap NLI4CT-P aims to fill, through the application of targeted interventions and the introduction of novel evaluation metrics.

## 7 Conclusion

This study introduces the NLI4CT-P dataset and provides a comprehensive analysis of submissions to SemEval-2024 Task 2, underscoring the persistent challenges in Clinical Natural Language Inference (NLI) despite significant advancements in Large Language Models (LLMs). The incorporation of Faithfulness and Consistency metrics further highlights these challenges, shedding light on areas requiring additional focus, if these systems are to meet the requirements for real-world clinical implementation. Our key findings reveal that generative models markedly outperform discriminative models, particularly in terms of Faithfulness and Consistency. The utility of additional data is underscored, especially due to the limited size of the NLI4CT-P training set. Furthermore, our analysis reveals the substantial impact of prompting strategies on model performance, noting an intriguing preference for zero-shot approaches over few-shot methods. Additionally, mid-sized architectures, ranging between 7B and 70B parameters, demonstrate the potential to match or even exceed the performance of larger models (>70B) in F1 scores, Faithfulness, and Consistency, while being more resource and cost-effective. Conversely, models with fewer than 7B parameters face difficulties in achieving comparable results. We plan to perform a further analysis of the submitted systems' performance at an intervention level, identifying specific areas of weakness, such as numerical reasoning or handling longer premises, to refine and enhance Clinical NLI systems further.

## 8 Limitations

Despite not disclosing detailed specifics of the interventions, nor providing intervened training data, several participants generated their own interventions for data augmentation. As a result, some models were specifically trained on this intervened data. However, this approach raises concerns regarding their ability to generalize effectively to entirely new, unseen perturbations or adversarial datasets. The tailored training to specific interventions may limit the models' broader applicability and robustness on unseen perturbed or adversarial data.

## 9 Acknowledgments

This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617), by the EPSRC grant EP/T026995/1 entitled "EnnCore: End-to-End Conceptual Guarding of Neural Architectures" under Security for all in an AI enabled society, by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre.

## References

Reem Abdel-Salam, Mary Adetutu Adewunmi, and Mercy Akinwale. 2024. Caresai at semeval-2024 task 2: Improving natural language inference in clinical trial data using model ensemble and data explanation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1916–1922, Mexico City, Mexico. Association for Computational Linguistics.

Mathilde Aguiar, Pierre Zweigenbaum, and Nona Naderi. 2024. Seme at semeval-2024 task 2: Comparing masked and generative language models on

natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 975–985, Mexico City, Mexico. Association for Computational Linguistics.

Abbas Akkasi, Adnan Khan, Mai A. Shaaban, Majid Komeili, and Mohammad Yaqub. 2024. iml at semeval-2024 task 2: Safe biomedical natural language inference for clinical trials with llm based ensemble inferencing. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 170–174, Mexico City, Mexico. Association for Computational Linguistics.

Duygu ALTINOK. 2024. D-nlp at semeval-2024 task 2: Evaluating clinical inference capabilities of large language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 600–614, Mexico City, Mexico. 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.

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*.

Greg Brockman, Atty Eleti, Elie Georges, Joanne Jang, Logan Kilpatrick, Rachel Lim, Luke Miller, and Michelle Pokrass. 2023. ChatGPT and Whisper APIs. `https://openai.com/api/`. Accessed: April 3, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Clement Brutti-Mairesse. 2024. Crcl at semeval-2024 task 2: Simple prompt optimizations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 424–429, Mexico City, Mexico. Association for Computational Linguistics.

Abir Chakraborty. 2024. Rgat at semeval-2024 task 2: Biomedical natural language inference using graph attention network. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 116–122, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jiaxu Dao, Zhuoying Li, Xiuzhong Tang, Xiaoli Lan, and Junde Wang. 2024. Puer at semeval-2024 task 2: A biolinkbert approach to biomedical natural language inference. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 70–75, Mexico City, Mexico. Association for Computational Linguistics.

Spandan Das, Vinay Samuel, and Shahriar Noroozizadeh. 2024. Tldr at semeval-2024 task 2: T5-generated clinical-language summaries for deberta report analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 507–516, Mexico City, Mexico. 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.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Anass Fahfouh, Abdessamad Benlahbib, Jamal Riffi, and Hamid Tairi. 2024. Usmba-nlp at semeval-2024 task 2: Safe biomedical natural language inference for clinical trials using bert. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 419–423, Mexico City, Mexico. 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.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Aryo Gema, Giwon Hong, Pasquale Minervini, Luke Daines, and Beatrice Alex. 2024. Edinburgh clinical nlp at semeval-2024 task 2: Fine-tune your model unless you have access to gpt-4. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1905–1915, Mexico City, Mexico. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

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.

Artur Guimarães, Bruno Martins, and João Magalhães. 2024. Lisbon computational linguists at semeval-2024 task 2: Using a mistral-7b model and data augmentation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1270–1277, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

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*.

Hwanmun Kim, Kamal raj Kanakarajan, and Malaikannan Sankarasubbu. 2024. Saama technologies at semeval-2024 task 2: Three-module system for nli4ct enhanced by llm-generated intermediate labels. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1423–1445, Mexico City, Mexico. Association for Computational Linguistics.

Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):1–17.

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.

Lung-Hao Lee, Chen-Ya Chiou, and Tzu-Mi Lin. 2024. Nycu-nlp at semeval-2024 task 2: Aggregating large language models in biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1465–1472, Mexico City, Mexico. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context-investigating contextual reasoning over long texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13388–13396.

Jin Liu and Steffen Thoma. 2024. Fzi-wim at semeval-2024 task 2: Self-consistent cot for complex nli in biomedical domain. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1259–1269, Mexico City, Mexico. Association for Computational Linguistics.

Yu-An Lu and Hung-Yu Kao. 2024. 0x.yuan at semeval-2024 task 2: Agents debating can reach consensus and produce better outcomes in medical nli task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 305–310, Mexico City, Mexico. Association for Computational Linguistics.

Shreyasi Mandal and Ashutosh Modi. 2024. Iitk at semeval-2024 task 2: Exploring the capabilities of llms for safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1386–1393, Mexico City, Mexico. Association for Computational Linguistics.

Jennifer Marks, MohammadReza Davari, and Leila Kosseim. 2024. Clac at semeval-2024 task 2: Faithful clinical trial inference. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1683–1687, Mexico City, Mexico. Association for Computational Linguistics.

Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023. A symbolic framework for systematic evaluation of mathematical reasoning with transformers. *arXiv preprint arXiv:2305.12563*.

Marius Micluța-Câmpeanu, Claudiu Creanga, Ana-Maria Bucur, Ana Sabina Uban, and Liviu P. Dinu. 2024. Unibuc at semeval-2024 task 2: Tailored prompting with solar for clinical nli. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 573–582, Mexico City, Mexico. Association for Computational Linguistics.

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

Shantanu Nath and Ahnaf Mozib Samin. 2024. Bd-nlp at semeval-2024 task 2: Investigating generative and discriminative models for clinical inference with knowledge augmentation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1291–1297, Mexico City, Mexico. 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.

Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 667–676.

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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

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

Julia Rozanova, Marco Valentino, and Andre Freitas. 2023. Estimating the causal effects of natural logic features in neural nli models. *arXiv preprint arXiv:2305.08572*.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. *arXiv preprint arXiv:2010.06060*.

Marco Siino. 2024. T5-medical at semeval-2024 task 2: Using t5 medical embedding for natural language inference on clinical trial data. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 40–46, Mexico City, Mexico. Association for Computational Linguistics.

Veronika Smilga and Hazem Alabiad. 2024. Tüduo at semeval-2024 task 2: Flan-t5 and data augmentation for biomedical nli. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 723–730, Mexico City, Mexico. Association for Computational Linguistics.

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

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2022. A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint arXiv:2210.12023*.

Zihang Sun, Danqi Yan, Anyi Wang, Tanalp Agustoslu, Qi Feng, Chengzhi Hu, Longfei Zuo, Shijia Zhou, Hermine Kleiner, Pingjun Hong, Suteera Seeha, Sebastian Loftus, Anna Barwig, Oliver Kraus, Jona Volohonsky, Yang Sun, Leopold Martin, Lena Altinger, Jing Wang, and Leon Weber. 2024. Lmu-bionlp at semeval-2024 task 2: Large diverse ensembles for robust clinical nli. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1587–1593, Mexico City, Mexico. Association for Computational Linguistics.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. *arXiv preprint arXiv:1804.08117*.

Bhuvanesh Verma and Lisa Raithel. 2024. Dfki-nlp at semeval-2024 task 2: Towards robust llms using data perturbations and minmax training. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 668–682, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021. Measure and improve robustness in nlp models: A survey. *arXiv preprint arXiv:2112.08313*.

Yuqi Wang, Zeqiang Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. 2024. Dke-research at semeval-2024 task 2: Incorporating data augmentation with generative models and biomedical knowledge to enhance inference robustness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 88–94, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. *arXiv preprint arXiv:2009.07964*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sicheng YU, Jing JIANG, Hao ZHANG, Yulei NIU, Qianru SUN, and Lidong BING. 2022. Interventional training for out-of-distribution natural language understanding.

Rengui Zhang, Jin Wang, and Xuejie Zhang. 2024. Ynuhpcc at semeval-2024 task 2: Applying deberta-v3-large to safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 772–778, Mexico City, Mexico. 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*.

# A Intervention Prompts

## A.1 Contradictory Rephrasing prompt

*Your task is to provide 3 contradictory statements, given an original statement.*

*(Instructions) Ensure that the contradictory statements are factually opposed to the original statement. Do not mention the original statement in the contradictory statements. Use formal and straightforward language when writing the new statements, and avoid unusual or overly descriptive language. Make sure to retain the names 'Primary Clinical Trial' and 'Secondary Clinical Trial' in the contradictory statements, these names must be present in every statement. Provide 3 different options in a consistent JSON format with keys 'Statement_1', 'Statement_2', and 'Statement_3' followed by their respective paraphrased statements.*

*(Examples) 1. [original statement]:"the secondary trial requires patients to be over a certain age, but the primary trial does not specify an age range for participation." [ideal output]: "the secondary trial does not give an age limit for patients to participate, but patients must be between the age of 12-34 to be eligible for the primary trial"*

*2. [original statement]:"a patient that has received an organ transplant within the last month, and is still bedridden would be excluded from the primary trial but may be eligible for the secondary trial" [ideal output]: "a patient that has received an liver transplant in the last week, with an ECOG score of 4 would be eligible for the primary trial but excluded from the secondary trial"*

*3.[original statement]: "Women with Newly diagnosed stage IV breast cancer, confirmed as ER+ Considering a mastectomy are eligible for the primary trial" [ideal output]: "Women recently diagnosed with stage 4 ER-positive breast cancer and contemplating a mastectomy are excluded from the Primary Clinical Trial"*

*Input:*

### A.2 Paraphrasing prompt

*Your task is to provide 3 paraphrased statements, given an original statement.*

*(Instructions) Use formal and straighforward language when writing the new statements, and avoiding unusual or overly descriptive language. Make sure to retain the name 'Primary Clinical Trial' in the statements, this name must be present in every statement. Provide 3 different options in a consistent JSON format with keys 'Statement_1', 'Statement_2', and 'Statement_3' followed by their respective paraphrased statements.*

*(Examples) 1. [original statement]:"the primary trial does not specify an age range for participation." [ideal output]: "patients aged between 30-60 years old can be eligible for the primary trial"*

*2. [original statement]:"a patient that has received an organ transplant within the last month, and is still bedridden would be excluded from the primary trial" [ideal output]: "a patient that has received an liver transplant in the last week, with an ECOG score of 4 would be excluded from the primary trial"*

*3.[original statement]: "Women with Newly diagnosed stage IV breast cancer, confirmed as ER+ Considering a mastectomy are eligible for the primary trial" [ideal output]: "Women recently diagnosed with stage 4 ER-positive breast cancer and contemplating a mastectomy are suitable for the Primary Clinical Trial"*

*Input:*

### A.3 Numerical Paraphrasing prompt

*Your task is to modify the numerical values and units in an original statement while maintaining its original meaning, to generate 3 new statements.*

*(Instructions) Do not paraphrase the statements, You can only change numerical values or units, if you change the units you must also convert the measurement values. Provide 3 different options in a consistent JSON format with keys 'Statement_1', 'Statement_2', and 'Statement_3' followed by their respective paraphrased statements.*

*(Examples) 1. [original statement]: "Over 6 weeks of TAK-228 Plus Tamoxifen treatment patients in the primary trial experienced a 5% reduction in the Percentage of cells with Ki67 expression" [ideal output]: "Over 42 days of TAK-228 Plus Tamoxifen treatment patients in the primary trial experienced a 5% reduction in the Percentage of cells with Ki67 expression"*

*2.[original statement]: "in the primary trial there were 10 times the number of Hepatotoxicity cases as there were cases of hypertension and Pancreatectomy" [ideal output]: "in the primary trial there were 1000% the number of Hepatotoxicity cases as there were cases of hypertension and Pancreatectomy"*

*3.[original statement]: "2/73 the primary trial participants, and 0/1674 the secondary trial participants suffered an Acute myocardial infarction " [ideal output]: "2.74% the primary trial participants, and 0% the secondary trial participants suffered an Acute myocardial infarction "*

*Input:*

### A.4 Numerical Contradictory Rephrasing prompt

*Your task is to modify the numerical values and units in an original statement to contradict the original statement, to generate 3 new statements.*

*(Instructions) Do not paraphrase the statements, You can only change numerical values or units, if you change the units you must also convert the measurement values. Provide 3 different options in a consistent JSON format with keys 'Statement_1', 'Statement_2', and 'Statement_3' followed by their respective paraphrased statements.*

*(Examples) 1. [original statement]: "Over 6 weeks of TAK-228 Plus Tamoxifen treatment patients in the primary trial experienced a 5% reduction in the Percentage of cells with Ki67 expression"*

*[ideal output]: "Over 50 days of TAK-228 Plus Tamoxifen treatment patients in the primary trial experienced a 105% reduction in the Percentage of cells with Ki67 expression"*

*2.[original statement]: "in the primary trial there were 10 times the number of Hepatotoxicity cases as there were cases of hypertension and Pancreatectomy" [ideal output]: "in the primary trial there were 30% the number of Hepatotoxicity cases as there were cases of hypertension and Pancreatectomy"*

*3.[original statement]: "2/73 the primary trial participants, and 0/1674 the secondary trial participants suffered an Acute myocardial infarction " [ideal output]: "9.74% the primary trial participants, and 8% the secondary trial participants suffered an Acute myocardial infarction "*

*Input:*