A Reproducibility Study on Consistent LLM Reasoning for Natural Language Inference over Clinical Trials

Artur Guimarães[1] [0000−0003−3682−4960],
João Magalhães[2,3] [0000−0001−6290−5719], and
Bruno Martins[1] [0000−0002−3856−2936]

[1] INESC-ID
[2] NOVA-LINCS
[3] NOVA University of Lisbon
{artur.guimas@gmail.com, jmag@fct.unl.pt,
bruno.g.martins@tecnico.ulisboa.pt}

Abstract. With the rapid expansion of AI in healthcare, ensuring that language models can reason accurately and consistently within the medical domain is essential for enhancing clinical decision-making. Consistent reasoning is particularly challenging, since once a model outputs a judgment for a given medical statement, it should retain that judgment when faced with a syntactically altered version of the statement. In addition, when the altered version of the statement reflects a semantic shift, the model should adjust its judgment accordingly. In this paper, we describe the process of reproducing state-of-the-art methods for safe biomedical Natural Language Inference for Clinical Trials (NLI4CT), emphasizing model vulnerability to small input variations and the inherent complexity of clinical trial reports (CTRs). We specifically evaluate the reasoning capabilities of Large Language Models (LLMs) in the SemEval-2024 NLI4CT dataset, thus considering a task that focused on robustness and which serves as a good proxy for real-world applications. To improve thoroughness, we extended this study and explore a broader set of techniques, establishing baseline scores for several widely used models. We conclude with an analysis of the results, highlighting key insights and empirical lessons that contribute to future research in this domain.

Keywords: Natural Language Inference · Reasoning with LLMs · Clinical Trials · SemEval-2024 · Reproducibility.

1 Introduction

Clinical Trials (CTs) [15,39] are research studies that aim to introduce new medical treatments safely [2]. These studies are documented in Clinical Trial Reports (CTRs), which provide concise descriptions of the objectives of a trial, its methodologies, eligibility criteria, and outcomes. Clinical experts can for instance rely on CTRs to evaluate the suitability of administering new treatments to patients, which is a complex reasoning task.

Section ID: Interventions
Primary Clinical Trial:
INTERVENTION 1: PF-06647020 0.2 mg/kg (Q8Q3W Regimen); Participants enrolled in the dose escalation phase received PF-06647020 at 0.2 mg/kg on Day 1 of each 21-day cycle ...
INTERVENTION 2: PF-06647020 0.5 mg/kg (Q3W Regimen); Participants enrolled in the dose escalation phase received PF-06647020 at 0.5 mg/kg on Day 1 of each 21-day cycle ...
Secondary Clinical Trial:
INTERVENTION 1: ALT-801 0.015 mg/kg/Dose;
INTERVENTION 2: ALT-801 0.040 mg/kg/Dose
Statement: The drug dosage for the primary clinical trial participants is lower in comparison to that of the secondary clinical trial participants.

Fig. 1. Example of an entailment/contradiction query in the clinical domain.

Large Language Models (LLMs) offer significant potential in assisting clinical reasoning by interpreting and deriving conclusions from CTRs, among other use cases. Nonetheless, several critical challenges limit their reliability, including data contamination [24,7,11], fairness concerns [32,22], generation of hallucinations [40,1], or lack of explainability [6,41]. Most importantly for this work, the nonexistence of robust evaluation metrics specifically designed to assess a model's reasoning ability, combined with limited comprehensive benchmarks [28,8,27] for CTRs, conjointly pose significant concerns in ensuring safe and reliable systems.

To enhance evaluation methodologies for clinical Natural Language Inference (NLI), SemEval-2024 [25] organized the Safe Biomedical Natural Language Inference for Clinical Trials (NLI4CT) task [16]. This task involved the classification of statements related to CTRs, establishing a framework for the adequate assessment of LLMs in the context of clinical reasoning. More precisely, the primary objective was to support the development of systems capable of predicting whether a given statement can be entailed or contradicted by the information contained within one or two CTRs, in a consistent and coherent manner.

Considering that the same CTR can be associated to different variations of a base statement, systems participating in the NLI4CT task were required to demonstrate consistent responses to these modifications. Desired system properties included accuracy and robustness in the face of redundant perturbations to the statements, and an alignment between semantic changes in the input and the system's output. The idea is to ensure that systems exhibit known and expected error modes, this way reinforcing reliability and interpretability.

As the main contribution of this paper, we reproduce and publish the implementations of state-of-the-art NLI methods in the clinical domain. This effort offers as a rigorous and reproducible empirical study designed to advance clinical reasoning research. In addition to reproducing previous methods, we conducted a series of experiments that extend current methodologies, yielding notably strong results that not only reaffirm existing baseline performances, but also provide new insights into their efficacy.

2 Related Work

This section contextualizes the available resources for clinical NLI, and provides a summarized overview of the main techniques that performed well in the SemEval-2024 NLI4CT challenge.

Julien et al. [17] provided an exposition on benchmark datasets available in the domain of clinical trials: the data from the **TREC 2021 Clinical Track** [29,27], focused on associating synthetic patient cases to CTRs by retrieving the most relevant results to each patient, evaluated by nDCG and precision; The **MEDNLI** [28] dataset, that focuses entailment relations between patient medical records and annotated sentences about those records, evaluating accuracy; and **Evidence Inference 2.0** [8], in which systems have to perform Question Answering (QA) with the goal of predicting the relationship between given CTR data and a treatment, evaluated by macro F1, AUC, and an evidence token metric. The NLI4CT dataset [17] distinguishes itself from the aforementioned datasets by utilizing full CTR sections in its queries, by containing examples that intend to prove a model's ability to perform numerical inference and common-sense reasoning, and by focusing on ensuring robustness to perturbations on statements.

NLI4CT is a particularly important benchmark dataset in the clinical domain because it specifically targets the LLMs reasoning capacities. The following teams submitted interesting approaches and achieved top scores in the SemEval-2024 NLI4CT challenge [16]: Gema et al. [10] (Edinburgh Clinical NLP); Lee et al. [21] (NYCU-NLP); Liu and Thoma [23] (FZI-WIM); and Guimarães et al. [12] (Lisbon Computational Linguists). A common starting point was to assess the **zero-shot** performance of several LLMs, without providing examples or using a refined prompting strategy. The standout result from this process was obtained by utilizing GPT-4 [26], which outperformed all other approaches, whilst other experiments served as a baseline result to ablate the performance gains from training procedures or advanced prompting techniques.

Previous studies demonstrated the effects of **prompting** [5] on the capability that a LLM has to follow instructions, exposing emergent abilities from the pre-training process [4,37]. To this effect, teams explored two main prompting strategies: **Chain-of-Thought (CoT)** reasoning [38], instructing the model to explain its sequential inference process with specific key phrases (e.g., "*let's think step by step*" or "*verify it step by step*") [19]; and **In-Context Learning (ICL)** [34], where examples of one or more successful answers to CTR-Statement pairs were added to the prompt. Finally, **Self-Consistency** was also tested in conjunction with the aforementioned techniques [35], taking advantage of the ability to generate non-deterministic outputs (e.g., using sampling with a given temperature, top-p and top-k parameters) to explore different reasoning chains, and choosing the final label by majority voting.

Advancing to techniques that require training, all the aforementioned submissions to the NLI4CT task explored using **instruction fine-tuning** [36] on open-source LLMs to improve their reasoning abilities, testing mostly models based on Mistral-7B [14] and LLaMa-2 [31]. To increase the amount and quality of fine-tuning data, some authors **augmented the available data**, using

Original Statement: Only 2 patients in the primary trial did not have Recurrence-free Survival (Label: `Entailment`)

(a) Variation–Paraphrase: Only two individuals in the primary clinical trial were not recurrence-free survivors (Label: `Entailment`)

(b) Variation–Contradiction: Adverse events were a common occurrence among participants in the primary trial (Label: `Contradiction`)

(c) Variation–Text Append: Aggravated malignant neoplasm is a malignant neoplasm that shows clinical and/or pathologic progression. Only 2 patients in the primary trial did not have Recurrence-free Survival. (Label: `Entailment`)

(e) Variation–Num. Contradiction: 3% of patients in the primary trial did not experience any adverse events. (Label: `Contradiction`)

Fig. 2. Examples of variations on a clinical statement.

Chat-GPT v3.5 [18], GPT-4, or Mistral-7B to either explain training examples (e.g., using CoT prompts [13]) or to create examples from the original statements (e.g., paraphrasing, contradicting or elaborating on the original statements).

3 Aspects Involved on Consistent Reasoning Methods

This section highlights the key challenges of the task at hand, including the complexity of consistent reasoning and the nuanced nature of CTRs. It then introduces the state-of-the-art methods that were implemented for this study, detailing the rationale behind each model choice.

3.1 Consistent Reasoning about Clinical Trials

Building on the previously established challenges, the core concept behind consistent reasoning is to design systems that arrive at their conclusions based on valid reasoning processes, even if their final responses are incorrect. This is crucial because, without insight into the specific factors driving a system's judgment, it becomes difficult to assess its true reliability. Therefore, it is essential to develop evaluation methodologies that provide a deeper understanding of how dependable the systems are.

Figure 2 illustrates this process with a practical example. Consider the original statement "*only two patients in the primary trial did not have recurrence-free survival*," analyzed within the context of a CTR. The correct classification for the example Statement-CTR pair is "*entailment*". If we were to proceed without further assessment, we would have little assurance that the system truly understood and analyzed the example, or if it merely reached the conclusion through reasoning shortcuts. To address this uncertainty, we introduce variations of the statement to test the model's robustness: (a) simple paraphrasing examines whether the system can maintain its judgment despite syntactic changes; (b) contradiction tests its ability to adjust the label in response to semantic shifts; (c) appending text introduces irrelevant noise; (d) numerical paraphrasing alters

quantities whilst keeping meaning (e.g., changing "*2/67 patients*" to " *3%*"); and (e) numerical contradiction changes quantities and their meaning. Having established the importance of consistent reasoning and the need for robust evaluation methodologies, we now turn our attention to method selection.

3.2 Baseline Methods based on LLMs

After reviewing the average scores of the twenty-five systems that were evaluated in the SemEval-2024 NLI4CT task [16], we focused on the success of the top methods discussed in Section 2. Before reproducing their work, we elected to establish a comprehensive and unbiased analysis of **zero-shot** results, as a baseline for all other techniques. We opted for a selection of **LLMs** that allowed us to reproduce (not replicate) the top submissions, namely the following:

1. **Mistral-7B** based models: **Mistral-7B-instruct-v0.2**[4], fine-tuned with instruction prompts; **BioMistral-7B** [20][5], pre-trained with medical data; and **MistralLite-7B**[6], fine-tuned with longer text sequences.
2. **LLaMa3.1-8B-instruct** [9][7], a recent version of the LLaMa herd of models;
3. **Gemini** [30], a commercial black-box model chosen to emulate GPT-4.

As we focused on **zero-shot** performance only, we did not use ICL, CoT, or self-consistency in our baselines. Each model was **prompted** with both a **short** and a **long prompt**[8], where the short prompt provided minimal context and the long prompt detailed the structure of a CTR. The prompts were adapted to suit each model's templating requirements (e.g., correct use of special tokens). For **decoding strategies**, we tested both **greedy** decoding and **sampling**.

The purpose of establishing baselines is twofold: first, we aimed to evaluate each model's core capabilities without applying any additional techniques, ensuring an accurate assessment of their performance, as prior studies have shown varying results with the same models. Second, we intend for our techniques to enhance these baseline outcomes, which requires rigorous comparison across multiple iterations, to measure the improvements effectively. To ensure the reproducibility and robustness of our results, we repeated the tests for each sampling model configuration using five random seeds, reporting the **maximum average score** from these runs, as in the original task leaderboard teams were allowed to submit several runs and kept the highest score from them. While we acknowledge that small adjustments, particularly to the prompts, could significantly impact the results, we believe this approach allows for a valid comparative analysis.

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[5] https://huggingface.co/BioMistral/BioMistral-7B
[6] https://huggingface.co/amazon/MistralLite
[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[8] https://anonymous.4open.science/r/ECIR-2025_Reproducibility_NLI4CT-8D58/src/prompts/README.md

3.3 Reproduced Clinical Reasoning Methods

Due to the breadth of models and techniques across multiple systems, and the iterative nature of each approach, reproducing every preliminary experiment in exhaustive detail would be impractical. Instead, we summarize the core contributions of state-of-the-art methods, presenting their respective scores in Tables 3 and 4. Given the primary goal of reproducing, rather than replicating, we applied small but impactful modifications to each system, resulting in significant performance differences.

Gema et al. [10] employed frozen models to evaluate various settings: ICL (1-shot and 2-shot), CoT, and combined approaches (CoT + 1-shot, CoT + 2-shot), with reasoning chains generated via Chat-GPT. They also incorporated Parameter-Efficient Fine-Tuning (PEFT) with zero-shot prompting. In reproducing their work, we applied the following adjustments: we replaced their ICL + CoT reasoning chains with those extracted by Liu and Thoma[23] from GPT-4, utilized a different method for response label extraction, adjusted the PEFT hyperparameters slightly, and selected alternative models (switching from LLaMa2-13B to LLaMa3.1-8B-instruct and from GPT-4 to Gemini). The change from GPT-4 to Gemini was imposed by budget constraints, as we had Gemini availability through our research projects and did not possess infrastructure to support utilizing bigger open-source LLMs (e.g., Mixtral-8x7B Instruct or LLaMa-3.2-70B), which would be our preferred choice to ensure reproducibility.

For Lee et al.[21], PEFT was employed to train on a synthetic set of data augmented by Chat-GPT 3.5. Additionally, Open-Chat[33] was used to revert variations of statements to their original forms before processing. Since this work did not provide open-source code, we focused instead on assessing the performance of other experimental setups.

In the work by Liu and Thoma [23], PEFT was applied to fine-tune Mixtral-8x7B on data synthetically augmented by GPT-4, generated through CoT reasoning on the NLI4CT train set. Once the model was fine-tuned, they assessed its zero-shot, CoT, and CoT + self-consistency performance. For our reproduction, we made several changes: we used a smaller model (Mistral-7B-instruct-v0.2) to accommodate our limited resource availability, implemented a different response label extraction method, altered prompt syntax to align with Guimarães et al. [12], and sampled five reasoning chains instead of their original ten.

Finally, Guimarães et al. [12] evaluated PEFT using manually and synthetically augmented data generated with Mistral-7B-instruct-v0.1. In reproducing their work, we opted for a more concise prompt structure, modified their PEFT hyper-parameters, utilized an alternative method for label extraction, and experimented with decoding strategies, including both greedy decoding and adjustments to sampling parameters. We did not incorporate any additional training data, using the original augmented data from Guimarães et al. [12].

3.4 Implementation Details

We implemented our work entirely in Python, relying primarily on the `torch`, `transformers`, and `huggingface` libraries. The `huggingface` platform was used to obtain all the model versions needed for the experiments.

To maintain consistency with prior studies, we adhered to the original prompts provided by each source, making adjustments when necessary to fit the templating requirements, or where we managed to improve performance. Instead of replicating the specific pre-processing techniques for CTRs, used by different studies, we employed the pre-processing functions from Guimarães et al. [12]. Additionally, we did not use the exact same model versions as in the original studies but opted for updated versions when available. Our goal was not strict replication but rather to reproduce the underlying ideas and methodologies, allowing for modifications in implementation to potentially enhance the results.

For the text generation processes, we handled greedy generation by simply setting the `Sample` flag to `False` in the generation. In contrast, for sampling generation, we applied two different configurations based on whether self-consistency was used. When self-consistency was applied, we set `temperature = 1.0`, `top-p = 0.99`, and `top-k = 50`. Otherwise, we adjusted these parameters to `temperature = 0.5`, `top-p = 0.7`, and `top-k = 15`.

To extract the correct label from the generated response, we first converted the entire output to lowercase and identified whether the terms associated with "entailment" or "contradiction" appeared first (e.g., "entailment," "entails," "entailed," etc., for entailment, and "contradiction," "contradicts," "contradicted," etc., for contradiction). We assigned the corresponding label unless a negation word (e.g., "not," "no," "isn't") was present immediately before these terms, in which case the label was reversed. Additionally, following the approach of Gema et al. [10], we reversed the response string for CoT reasoning, allowing us to search for the last occurrence of "Entailment" or "Contradiction". This method reduced label extraction errors, as the model's final tokens, where the conclusion is typically expressed, are critical for accurately capturing the intended response with the context of its own reasoning chain.

4 Experimental Methodology

This section describes the methodology employed to reproduce the experiments and results from the analyzed systems. We detail the key steps taken to ensure consistency with the original studies, while addressing any challenges that arose during reproduction.

4.1 Dataset

The analyzed CTRs were sourced exclusively from the United States Library of Medicine[9], comprising 1,000 English-language trial reports focused on breast

[9] https://clinicaltrials.gov/

Table 1. Full dataset composition for the SemEval-2024 NLI4CT Task

Split	#Pairs	Label (%)		Type (%)		Section (%)			
		Entail.	Contra.	Single	Comp.	Inter.	Elig.	Results	Adv.Eve.
Train	1700	50.0	50.0	61.0	39.0	23.3	28.6	18.9	29.2
Dev	200	50.0	50.0	70.0	30.0	18.0	28.0	28.0	26.0
Test	5500	33.5	66.5	46.4	53.6	28.0	25.8	22.5	23.7

cancer. Each CTR is structured into four sections: **Eligibility Criteria**, **Interventions**, **Results**, and **Adverse Events**. The task narrows its focus by having each clinical statement pertain to one of these sections. An example of a statement referencing two CTRs is shown in Figure 1.

Ground-truth labels for each Statement-CTR(s) pair are assigned as either **entailment**, if the CTR(s) fully supports the statement, or **contradiction**, if the statement is contradicted or lacks supporting information in the CTR(s). Additionally, each instance is labeled based on whether it refers to one or two CTRs. **Single** instances involve a direct comparison between the statement and a single CTR, whereas **comparison** instances may require cross-referencing between two CTRs. NLI4CT is designed to reflect practical applications, serving as a foundation for the real-world deployment of models.

The dataset is divided into **training**, **development**, and **testing** splits, as detailed in Table 1. The test set includes 5,500 statements, divided into 500 original and 5,000 contrast statements. Contrast statements are variations of the original ones, generated through processes such as paraphrasing, text appending, or contradiction. These variations are deliberately crafted to potentially change the original labels.

4.2 Metrics

Through evaluation metrics, the task aimed to emphasize not only the accuracy of label predictions. but also the resilience of these predictions to variations in the original statements (an example is presented on Figure 2). To achieve this, the standard metrics were the following[10]:

- **Macro F1-Score**, serving as a baseline performance metric (not measuring robustness), and which is calculated by the arithmetic mean of precision and recall, averaged over both possible classes;
- **Faithfulness**, measuring the ability of a system to arrive at the correct prediction for the correct reason, which is estimated by calculating the ability of a model to change its label prediction when a statement undergoes a **semantic-altering** process. For N test set queries x_i within the contrast set C that corresponds to variations of statement y_i, and with f as the

[10] https://sites.google.com/view/nli4ct/semeval-2024/evaluation

model's output for each statement, we define faithfulness as:

$$\text{Faithfulness}\,(N) = \frac{1}{|N|}\sum_{1}^{N} \mid \text{f}\,(y_i) - \text{f}\,(x_i) \mid \tag{1}$$

where $x_i \in C$, $\text{Label}\,(x_i) \neq \text{Label}\,(y_i)$ and $\text{Label}\,(y_i) = \text{f}\,(y_i)$.

– **Consistency**, measuring the ability of a system to predict the same label for equivalent statements, which is estimated by calculating the ability of a model to maintain its prediction when a statement undergoes a **semantic-preserving** process. For N test set queries x_i within the contrast set C that corresponds to variations of statement y_i, and with f as the model's output for each statement, we define consistency as:

$$\text{Consistency}\,(N) = \frac{1}{|N|}\sum_{1}^{N} 1 - \mid \text{f}\,(y_i) - \text{f}\,(x_i) \mid \tag{2}$$

where $x_i \in C : \text{Label}\,(x_i) = \text{Label}\,(y_i)$ and $\text{Label}\,(y_i) = \text{f}\,(y_i)$.

5 Experimental Results and Discussion

5.1 Baseline Results

Table 2 presents the results for each model discussed in Subsection 3.2, with different prompt types and text generation methods. Overall, **Mistral-7B** performs well on short prompts, achieving an average score of 69.5 regardless of the text generation method used. However, its performance drops on long prompts, with averages of 63.1 for greedy generation and 63.0 for sampling. In contrast, **BioMistral-7B** delivers significantly weaker results, particularly with long prompts. While the short prompt results are moderate, with averages of 57.5 and 55.8, its performance on long prompts is highly inconsistent, marked by poor F1-scores of 2.4 and 39.3. The model's high faithfulness scores (99.7 and 80.3) can be attributed to it frequently outputting the label "contradiction," undermining faithfulness as a reliable metric, as the system fails to adjust to variations in the input.

Both **MistralLite-7B** and **LLaMa3.1** display more stable performance across prompt types and generation methods. **MistralLite-7B** maintains average scores around 61-62 for both prompt lengths, though its F1-scores for short prompts are notably weak. In contrast, **LLaMa3.1** consistently achieves the highest performance, with an average score reaching 72.5. Its results remain steady across both greedy and sampling techniques.

Among the open-source models we evaluated, **Mistral-7B** demonstrates a solid balance across F1-score, faithfulness, and consistency, while **LLaMa3.1** emerges as the overall strongest baseline. However, the commercial closed-source **Gemini** models set distinct performance benchmarks, excelling across all metrics. In particular, **Gemini-1.5-pro-001** achieves the highest average score, reaching 82.3 on short prompts with greedy decoding.

Table 2. Baseline scores by model, prompt type and decoding technique.

Model	Prompt	Decoding	Base Metrics			Average
			F1	Faithfulness	Consistency	
Mistral-7B-Instruct-v0.2	short	Greedy	72.4	70.9	65.2	69.5
Mistral-7B-Instruct-v0.2	short	Sampling	72.1	71.3	65.0	69.5
Mistral-7B-Instruct-v0.2	long	Greedy	70.6	56.6	62.0	63.1
Mistral-7B-Instruct-v0.2	long	Sampling	70.6	56.7	61.7	63.0
BioMistral-7B	short	Greedy	61.3	55.0	56.1	57.5
BioMistral-7B	short	Sampling	64.3	49.4	53.7	55.8
BioMistral-7B	long	Greedy	2.4	99.7*	61.6	54.6
BioMistral-7B	long	Sampling	39.3	80.3*	57.5	59.0
MistralLite-7B	short	Greedy	33.8	89.5	61.6	61.6
MistralLite-7B	short	Sampling	42.5	83.9	58.9	61.8
MistralLite-7B	long	Greedy	58.0	68.4	59.4	61.9
MistralLite-7B	long	Sampling	53.7	68.4	56.9	59.7
LLaMa3.1-8B-Instruct	short	Greedy	59.0	87.5	69.7	72.1
LLaMa3.1-8B-Instruct	short	Sampling	61.7	86.2	69.6	72.5
LLaMa3.1-8B-Instruct	long	Greedy	67.5	80.3	69.2	72.3
LLaMa3.1-8B-Instruct	long	Sampling	67.4	79.6	68.7	71.9
Gemini-1.5-flash-001	short	Greedy	75.5	86.3	75.2	79.0
Gemini-1.5-flash-001	long	Greedy	71.6	84.7	71.8	76.0
Gemini-1.5-pro-001	short	Greedy	**76.6**	**92.6**	**77.7**	**82.3**
Gemini-1.5-pro-001	long	Greedy	73.6	92.1	74.1	79.9

5.2 Experiments without Fine-Tuning

To reproduce the state-of-the-art results we started by exploring the notable work of Gema et al. [10] with frozen models, specifically with **Mistral-7B**, **LLaMa3.1**, and **Gemini**. Our evaluations spanned the same techniques explored in the original paper, including zero-shot, in-context examples (1-shot and 2-shot), chain-of-thought reasoning, and combinations of these methods. Table 3 presents a comparison between the original results and our reproduced results, utilizing the official task metrics.

An immediate observation is that we successfully improved consistency across all scenarios and saw gains in most other metrics across our experiments. The one exception was our GPT-4 reproduction, where budget constraints required the use of an alternative commercial model. Nevertheless, Gemini proved to be a robust baseline, and while Gema et al.'s scores with GPT-4 ranked highest on the leaderboard, our Gemini implementation would have secured a strong third-place position in the competition

We also observed significant disparities between the original and reproduced results, particularly in the non-CoT performance of **Mistral-7B**. The original study reported low faithfulness scores, which we substantially improved upon (13.4/70.9, 11.1/88.3, and 13.4/86.3). Additionally, when comparing the 0-shot performance of **Mistral-7B** with the baseline submitted by Guimarães et al. [12], we observed a modest but still notable improvement.

Table 3. Experiments with frozen models, marking *positive* or *negative* variations in terms of F1-Score, Faithfulness, Consistency and their Average measures.

Original Model / Our Model	Technique	Original Results/Our Results			
		F1	Faithfulness	Consistency	Average
Mistral-7B-Inst. [10]/ Mistral-7B-Inst.	0-shot	65.3/ 72.4	13.4/ 70.9	41.5/ 65.2	40.1/ 69.5
	1-shot	66.4/ 48.8	11.1/ 88.3	41.3/ 65.2	39.6/ 67.4
	2-shot	66.9/ 51.1	13.4/ 86.3	42.5/ 65.7	40.9/ 67.7
	CoT	47.1/ 48.2	59.3/ 79.6	50.8/ 60.5	52.4/ 62.8
	CoT+1-shot	58.4/ 57.4	57.1/ 74.0	54.9/ 62.1	56.8/ 64.5
	CoT+2-shot	59.4/ 59.5	60.7/ 77.5	56.5/ 63.7	58.9/ 66.9
LLaMA2-13B [10]/ LLaMa3.1-8B	0-shot	60.7/ 67.5	45.0/ 80.3	49.4/ 69.2	51.7/ 72.3
	1-shot	63.0/ 69.0	33.5/ 80.9	48.8/ 69.1	48.4/ 73.0
	2-shot	61.7/ 70.2	40.2/ 74.3	50.1/ 68.1	50.7/ 70.9
	CoT	60.3/ 64.5	50.1/ 88.2	51.2/ 68.3	53.9/ 73.7
	CoT+1-shot	63.5/ 69.4	53.1/ 84.7	53.6/ 71.3	56.7/ 75.1
	CoT+2-shot	59.2/ 68.5	61.2/ 87.0	55.5/ 72.8	58.6/ 76.1
Mistral-7B-Inst. [12]/ Mistral-7B-Inst.	0-shot	67.3/ 70.6	61.8/ 56.7	53.8/ 61.7	60.1/ 63.0
GPT-4 [10]/Gemini	0-shot	**77.5**/ 76.6	**94.8**/ 92.6	77.5/ **77.7**	**83.3**/ 82.3

5.3 Experiments with Fine-Tuning

Table 4 presents the results of our Parameter-Efficient Fine-Tuning (PEFT) setups, alongside the original findings. In our reproduction of Gema et al.[10], most metrics showed improvements with our modifications, though MistralLite-7B experienced a notable performance drop. In the case of Guimarães et al.[12], our adjustments significantly boosted faithfulness, though this came at the expense of F1-score and consistency, keeping average scores close to the original.

For Liu and Thoma's work [23], while our results were lower overall, it should be noted that the original model, Mixtral-8x7B, uses a mixture-of-experts configuration with 46.7B parameters—considerably larger than our Mistral-7B-Inst. Given this difference, our results are within the expected difference range, and even outperform expectations when adjusted for size disparities.

While we did not re-implement Lee et al. [21], their application of PEFT with variation reduction techniques shows robust performance, outperforming our reproduced models and ranking just below GPT-4.

5.4 Final Considerations

Overall, our reproduction efforts were largely successful in validating and even extending the original studies, albeit requiring additional discussion for the non-fine-tuned configurations. While we closely matched the original source methods in terms of generation configurations and overall prompt structures, there were notable differences in the implementation of key aspects. One such difference lies in how label outputs were extracted from the generated text. For instance,

Table 4. Models trained with **PEFT**, marking positive or negative variations in terms of F1-Score, Faithfulness, Consistency and their Average measures.

Original Model / Our Model	Technique	Original Results/Our Results			
		F1	Faithfulness	Consistency	Average
Mistral-7B-Inst. [10]/ Mistral-7B-Inst.	-	76.9/ 78.7	76.6/ 85.1	71.4/ 70.8	75.0/ 78.2
MistralLite-7B [10]/ MistralLite-7B	-	74.8/ 79.1	87.3/ 73.8	72.2/ 70.5	78.1/ 74.5
LLaMA2-13B [10]/ LLaMa3.1-8B	-	67.7/ 81.0	77.3/ 73.6	66.1/ 73.7	70.4/ 76.1
Mistral-7B-Inst. [12]/ Mistral-7B-Inst.	-	81.2/ 80.5	72.3/ 79.1	69.2/ 70.5	74.2/ 76.7
	Manual Aug.	82.9/ 77.6	76.9/ 83.4	71.9/ 69.9	77.2/ 77.0
	Synth. Aug.	78.1/ 77.8	78.0/ 85.6	71.0/ 70.8	75.7/ 78.1
	Mix Aug.	80.1/ 80.5	83.1/ 79.3	72.2/ 73.8	78.5/ 77.9
Mixtral-8x7B [23]/ Mistral-7B-Inst.	-	78.7/ 79.5	81.0/ 78.9	73.6/ 70.3	77.8/ 76.3
	CoT	78.7/ 73.9	89.7/ 84.7	72.2/ 69.7	80.2/ 76.1
	CoT+SC	80.0/ 79.0	90.4/ 84.0	72.9/ 69.1	81.1/ 77.4
Mistral-7B [21]/-	Reduction	76.2/———	86.1/———	78.1/———	80.1/———
Solar-10.7B [21]/-	Reduction	77.9/———	92.4/———	80.9/———	83.7/———

Gema et al. [10] identified the label by checking the last token for occurrences of "entailment" or "contradiction". Similarly, Liu and Thoma [23] employed a comparable approach but added terms like "inconclusive" or "undetermined" to classify uncertain outputs as contradictions. In contrast, Guimarães et al. [12] detected the label by checking the first occurrence of these tokens, which marked a significant departure from the method used in the present study.

Moreover, we deviated from the original pre-processing, resulting in discrepancies in how the CTRs were presented across different systems. Additionally, we adjusted the positioning of special tokens according to the specific templating guidelines of each system, which we found to have a significant impact, particularly in examples involving in-context examples. These variations highlight areas where even minor deviations in implementation can influence outcomes and provide insights for future reproducibility efforts.

6 Conclusions

This paper focused on reproducing and critically evaluating state-of-the-art methods for natural language inference on clinical trial reports, using the SemEval-2024 NLI4CT dataset as the evaluation setting. The key outcome is a robust and reproducible empirical study. We contributed with a rigorous replication of key approaches in the field, reinforcing the reliability of existing baselines while offering improvements in techniques that do not rely on fine-tuning. Furthermore, we successfully reproduced fine-tuning-based methods, achieving slight performance

gains across various configurations. By fulfilling the central objective of producing a comprehensive and reproducible analysis, this work lays the groundwork for future advancements. Our implementations have been made publicly available[11], in the hope that they will encourage continued research and refinement of these methods within the clinical domain.

For future directions, the benchmarking efforts outlined in this study offer a strong foundation for expanding this research to other datasets within the clinical domain, such as those from the TREC 2021 Clinical Track, MEDNLI, and Evidence Inference 2.0, could further extend its impact. Moreover, exploring the generalization of these techniques to other highly specialized fields, such as legal NLI [3], represents a promising avenue for further exploration.

Acknowledgments. This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), by Fundação para a Ciência e Tecnologia (FCT), through the project with reference UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020), and by the Google Cloud Platform Gift program, which provided funding to support model computation. We would also like to extend a thank you to the organizers that coordinated and all the teams that participated in the SemEval-2024 NLI4CT Task, especially the ones whose work was used in this paper and had the sympathy to publicly share their code.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in chatgpt: implications in scientific writing. Cureus **15**(2) (2023)
2. Avis, N.E., Smith, K.W., Link, C.L., Hortobagyi, G.N., Rivera, E.: Factors associated with participation in breast cancer treatment clinical trials. Journal of Clinical Oncology **24**(12), 1860–1867 (2006)
3. Bernsohn, D., Semo, G., Vazana, Y., Hayat, G., Hagag, B., Niklaus, J., Saha, R., Truskovskyi, K.: LegalLens: Leveraging LLMs for legal violation identification in unstructured text. In: Graham, Y., Purver, M. (eds.) Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2129–2145. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.eacl-long.130
4. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways (2022)
5. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv:2210.11416 (2022)
6. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable ai for natural language processing. arXiv:2010.00711 (2020)

[11] https://github.com/araag2/ECIR-2025_Reproducibility_NLI4CT

7. Deng, C., Zhao, Y., Tang, X., Gerstein, M., Cohan, A.: Investigating data contamination in modern benchmarks for large language models. arXiv:2311.09783 (2023)
8. DeYoung, J., Lehman, E., Nye, B., Marshall, I.J., Wallace, B.C.: Evidence inference 2.0: More data, better models. arXiv:2005.04177 (2020)
9. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv:2407.21783 (2024)
10. Gema, A., Hong, G., Minervini, P., Daines, L., Alex, B.: Edinburgh clinical NLP at SemEval-2024 task 2: Fine-tune your model unless you have access to GPT-4. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.) Proceedings of the International Workshop on Semantic Evaluation. pp. 1894–1904. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.semeval-1.265, https://aclanthology.org/2024.semeval-1.265
11. Golchin, S., Surdeanu, M.: Time travel in llms: Tracing data contamination in large language models. arXiv:2308.08493 (2023)
12. Guimarães, A., Martins, B., Magalhães, J.: Lisbon computational linguists at SemEval-2024 task 2: Using a mistral-7B model and data augmentation. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.) Proceedings of the International Workshop on Semantic Evaluation. pp. 1280–1287. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.semeval-1.185, https://aclanthology.org/2024.semeval-1.185
13. He, X., Wu, Y., Camburu, O.M., Minervini, P., Stenetorp, P.: Using natural language explanations to improve robustness of in-context learning for natural language inference. arXiv:2311.07556 (2023)
14. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv:2310.06825 (2023)
15. Jin, Q., Wang, Z., Floudas, C.S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., Lu, Z.: Matching patients to clinical trials with large language models. arXiv:2307.15051 (2023)
16. Jullien, M., Valentino, M., Freitas, A.: SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In: Proceedings of the International Workshop on Semantic Evaluation. Association for Computational Linguistics (2024)
17. Jullien, M., Valentino, M., Frost, H., O'Regan, P., Landers, D., Freitas, A.: Nli4ct: Multi-evidence natural language inference for clinical trial reports. arXiv:2305.03598 (2023)
18. Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., et al.: Chatgpt (large language model). https://chat.openai.com/chat (2023), https://chat.openai.com/chat
19. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
20. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv:2402.10373 (2024)
21. Lee, L.h., Chiou, C.y., Lin, T.m.: NYCU-NLP at SemEval-2024 task 2: Aggregating large language models in biomedical natural language inference for clinical

trials. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.) Proceedings of the International Workshop on Semantic Evaluation. pp. 1455–1462. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.semeval-1.209, https://aclanthology.org/2024.semeval-1.209

22. Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models. arXiv:2308.10149 (2023)
23. Liu, J., Thoma, S.: FZI-WIM at SemEval-2024 task 2: Self-consistent CoT for complex NLI in biomedical domain. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.) Proceedings of the International Workshop on Semantic Evaluation. pp. 1269–1279. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.semeval-1.184, https://aclanthology.org/2024.semeval-1.184
24. Magar, I., Schwartz, R.: Data contamination: From memorization to exploitation. arXiv:2203.08242 (2022)
25. Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.): Proceedings of the International Workshop on Semantic Evaluation. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.semeval-1.0
26. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., et al.: Gpt-4 technical report (2024)
27. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Bedrick, S., Hersh, W.R.: Overview of the trec 2022 clinical trials track. In: TREC (2022)
28. Romanov, A., Shivade, C.: Lessons from natural language inference in the clinical domain. arXiv:1808.06752 (2018)
29. Soboroff, I.: Overview of trec 2021. In: TREC (2021)
30. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv:2312.11805 (2023)
31. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
32. Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.W., Peng, N.: " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. arXiv:2310.09219 (2023)
33. Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., Liu, Y.: Openchat: Advancing open-source language models with mixed-quality data (2024)
34. Wang, X., Zhu, W., Saxon, M., Steyvers, M., Wang, W.Y.: Large language models are latent variable models: Explaining and finding good demonstrations for incontext learning. Advances in Neural Information Processing Systems **36** (2024)
35. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171 (2022)
36. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv:2109.01652 (2022)
37. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv:2206.07682 (2022)

38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
39. Wong, C., Zhang, S., Gu, Y., Moung, C., Abel, J., Usuyama, N., Weerasinghe, R., Piening, B., Naumann, T., Bifulco, C., et al.: Scaling clinical trial matching using large language models: a case study in oncology. In: Machine Learning for Healthcare Conference. pp. 846–862. PMLR (2023)
40. Xie, Q., Schenck, E.J., Yang, H.S., Chen, Y., Peng, Y., Wang, F.: Faithful ai in medicine: A systematic review with large language models and beyond. medRxiv (2023)
41. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology **15**(2), 1–38 (2024)