UniBuc at SemEval-2024 Task 2: Tailored Prompting with Solar for Clinical NLI

Marius Micluța-Câmpeanu[♣,♡,*], **Claudiu Creangă**[♣,♡,*]
Ana-Maria Bucur[♣,♡], **Ana Sabina Uban**[♠,♡], **Liviu P. Dinu**[♠,♡]
[♠] Faculty of Mathematics and Computer Science
[♣]Interdisciplinary School of Doctoral Studies, [♡]HLT Research Center
University of Bucharest, Romania
marius.micluta-campeanu@unibuc.ro, claudiu.creanga@s.unibuc.ro
ana-maria.bucur@drd.unibuc.ro, {auban, ldinu}@fmi.unibuc.ro

Abstract

This paper describes the approach of the UniBuc team in tackling the SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. We used SOLAR Instruct, without any fine-tuning, while focusing on input manipulation and tailored prompting. By customizing prompts for individual CTR sections, in both zero-shot and few-shots settings, we managed to achieve a consistency score of 0.72, ranking 14th in the leaderboard. Our thorough error analysis revealed that our model has a tendency to take shortcuts and rely on simple heuristics, especially when dealing with semantic-preserving changes.

1 Introduction

Clinical trials are prospective studies that aim to compare the effectiveness of an intervention against a control group in clinical patients (Friedman et al., 2015). ClinicalTrials.gov[1] hosts more than 480,000 clinical trials, making it challenging to analyze and extract information from them manually. Natural language inference has emerged as a valuable tool for interpreting evidence from clinical trials (Jullien et al., 2023a).

The second task of SemEval 2024 focuses on improving the understanding of clinical trial data through the second edition of NLI4CT (Natural Language Inference for Clinical Trials) (Jullien et al., 2024). This challenge is specifically designed to test the natural language inference capabilities of large language models (LLMs) and their ability to understand clinical text. The data used for the challenge was carefully annotated by clinical domain experts, and semantic interventions were performed to evaluate the safety and robustness of the models.

We employed LLMs, achieving the best results with SOLAR-Instruct, and focused on two key strategies:

- **Targeted Summarization**: Summarizing both CTRs and the hypothesis (retaining only the first 'trial' sentence) aided the model's focus on essential information.
- **Tailored Prompting**: We used both zero-shot and two-shot prompting, tailoring prompts to individual CTR sections for optimal results.

We were surprised to find that causal models significantly outperform masked language models on this type of task. This is probably because the task requires reasoning capabilities that BERT-based models do not have. Our model's biggest challenges were with numerical reasoning and rephrasing (discussed in Section 6), but despite those, we still secured the 14th place (out of 32) in the leaderboard. We make our code publicly available on GitHub[2].

2 Related Work

Recent work on clinical trial analysis includes detecting contradictions in medical publications (Makhervaks et al., 2023), automating eligibility assessment with LLMs (Wang et al., 2023; Datta et al., 2024), and assessing model hallucinations and reasoning capabilities in healthcare settings (Pal et al., 2023; Feng et al., 2023). The previous SemEval NLI4CT task (Jullien et al., 2023b) included a similar entailment task, with most submissions leveraging pre-trained language models. A small minority of approaches used ontologies and rule-based algorithms.

Few of the language model-based approaches include preprocessing of the data prior to using it as input to the models. In our approach for the current

[*]These authors contributed equally to this work.
[1]https://clinicaltrials.gov/
[2]https://github.com/ClaudiuCreanga/sem-eval-2024-task-2

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 586–595

task, we also attempt solutions based on both discriminative and generative language models, and in addition perform preprocessing of the input clinical trial data before feeding it to the models, such as summarization. While most of the best performing systems in last year's task employ some in-domain pre-training on medical data, we find that, from the models we experimented with, general LLMs perform as well as or better than medical ones. This could be explained by their larger size and instruction tuning techniques, but also because of the recent advances in general LLMs. However, we do not perform an exhaustive comparison of the two kinds of models (we use few medical LLMs in our experiments), so a definitive conclusion can not be drawn on this comparison.

The NLI4CT dataset (Jullien et al., 2023a) is a unique benchmark dataset for Natural Language Inference (NLI) in the clinical domain that contains data from Clinical Trial Reports (CTRs). In contrast, the MedNLI (Romanov and Shivade, 2018) dataset is another benchmark dataset for NLI in the clinical domain, but it only contains clinical notes. To ensure that NLI models are robust and safe, the organizers of the NLI4CT task perform semantic-preserving and semantic-altering interventions of the hypotheses. According to Jullien et al. (2023a), NLI models for clinical trials require not only biomedical reasoning but also numerical reasoning capabilities, as CTRs contain a large amount of quantitative information. In this regard, we conduct experiments using SOLAR 10B (Kim et al., 2023), which was trained on question-answer pairs from the mathematical domain to enhance its mathematical capabilities.

3 Data and Task Description

The data used for this task is comprised of 1,000 breast cancer CTRs collected from ClinicalTrials.gov with 24,000 entailment relations annotated by clinical domain experts (Jullien et al., 2023a). Each CTR is comprised of 4 sections: eligibility criteria, intervention, results and adverse events.

Each sample from the NLI4CT dataset consists of the CTR premise (one of the 4 sections of the CTR), a statement, and an entailment label (Entailment or Contradiction). The premise can refer to only one CTR in the *Single* type instance, or to a primary and a secondary trial in the *Comparison* type. The purpose of the current task is to evaluate the consistency of models and their ability to perform faithful reasoning (Jullien et al., 2024). For this purpose, different semantic altering (Contradiction Rephrasing and Numerical Contradiction) or semantic preserving interventions (Paraphrase, Numerical Paraphrase and Definition) have been conducted on the evaluation data. The NLI4CT dataset consists of 1,700 entailment relations in the training split, 200 in dev, and 5,500 in the test split.

To assess the performance of the models in the shared task, two metrics have been proposed: **Faithfulness** and **Consistency** (Jullien et al., 2024), besides F1-score. Faithfulness measures the model's ability to adjust its predictions accurately after a semantic-altering intervention. Consistency measures the capacity of the system to predict the same label for both the original and contrast statements, in the case of semantic-preserving interventions.

4 Methodology

In this section, we present the different approaches used to predict the entailment relations.

4.1 Pre-trained Masked Language Models

Our first approach for the task of entailment relations prediction is using pre-trained models. Previous research has shown that domain-specific pretraining is beneficial for biomedical tasks (Gu et al., 2022; Romanov and Shivade, 2018).

We use pre-trained models, such as PubMedBERT (Gu et al., 2021), XLM-RoBERTa (Conneau et al., 2020), Deberta V3 Large (He et al., 2021), and fine-tune them on the training data. The fine-tuning process is done in 2 steps, as suggested in (Sun et al., 2020). Firstly, we stack a fully connected layer on top of the pre-trained model and train it for 4 epochs while the weights of the pre-trained model are frozen with a learning rate of 10^{-3}. For the second step, we train the fully connected layer along with the last layer of the pre-trained model at a lower learning rate of $2 * 10^{-4}$ for one more epoch.

Inspired by the approach taken by Pahwa and Pahwa (2023), who demonstrated that the performance of fine-tuned cross-encoders is comparable to that of GPT-3.5 in the few-shot setting, we experiment with the sentence-transformer model BioBERT[3] trained on 6 benchmark NLI datasets (Deka et al., 2022) for sentence similarity tasks.

[3] pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

We train a cross-encoder model using the sentence embeddings from BioBERT on the NLI4CT train data for sentence-pair classification for 20 epochs.

We also utilized SciFive (Phan et al., 2021), a T5 model designed for biomedical literature-related tasks. This model was pre-trained on a large corpus of PubMed abstracts and PubMed Central full-text articles from biomedical and life sciences domains. It achieved state-of-the-art results on the MedNLI benchmark dataset (Romanov and Shivade, 2018). We used the SciFive model trained on MedNLI[4] in zero-shot setting to predict the entailment relations for the NLI4CT data.

4.2 Large Language Models

LLMs have achieved promising results in biomedical tasks, such as named entity recognition, relation extraction, text classification, and question answering (Jahan et al., 2023). We conducted experiments using LLaMa-2 7B[5] (Touvron et al., 2023), Mistral 7B[6] (Jiang et al., 2023) and SOLAR (Kim et al., 2023), which have shown promising results in various language tasks. LLaMa-2 is a competitive model that has performed well across multiple benchmarks such as commonsense reasoning, word knowledge, and reading comprehension. Mistral, on the other hand, has surpassed LLaMa-2 in all the tested benchmarks. We choose SOLAR-Instruct since it is a state-of-the-art model that is instruction-tuned specifically to have improved mathematical capabilities, with rephrased examples using a similar process to MetaMath (Yu et al., 2023). The team behind SOLAR developed a unique scaling technique, called depth up-scaling (DUS), which combines architectural changes with continued pre-training and obtained better results than larger models like Mixtral (Kim et al., 2023).

LLaMa-2 and Mistral were evaluated only in zero-shot settings on the NLI4CT test data. The same prompt was used in the experiments, regardless of the section the statement was referring to and it is presented in Appendix B. While the experiments from LLaMa-2 and Mistral were using the entire sections of CTRs and statements, we had a different approach for SOLAR, which involved CTRs summarization. We expand on the methodology below.

Section content summarization approach. This approach consists of two stages. First, we summarize the text of each section to reduce the number of tokens. Next, we perform the classification task on the shortened text using two-shot prompting with examples from the training set. Both stages of this pipeline use the SOLAR 10.7B Instruct v1.0 model (Kim et al., 2023) in GGUF format[7] with 5-bit quantization (Q5_K_M) using llama.cpp.

CTR summarization. We summarize the sections of each CTR in the evaluated datasets (train, development, test) to a maximum of 350 tokens. The main reason for performing summarization is to provide the model with a shorter context and a task that is easier to solve. To validate this statement, we perform preliminary runs on the development set on Single CTRs for the Results section. We obtain an F1-score of 0.55 on single results CTR without a summary, while we are able to reach 0.63-0.72 F1-score with the summarization approach.

Another motivating factor is the time required to run the inference in order to perform multiple experiments. Full CTR inference for one example can take up to 20-30 seconds. Conversely, shortened CTRs are evaluated in roughly 5 seconds. As generating summaries is a one-time cost, the time difference is compensated for after a few iterations. Evaluating on CTR summaries instead of the full CTR allowed us to dedicate more time in designing and refining prompts used for the entailment task. For these reasons, we do not conduct additional experiments on full CTRs.

The obvious drawback of summarization is the risk of discarding essential information. In the initial experiments for this approach, we tried to mitigate this by conditioning the summary to be related to the hypothesis statement. Unfortunately, this strategy caused the model to include the statement in the summary and at times even output contradictory phrases. Moreover, we did not try to continue with contextualized summaries because it would require generating a new summary for each statement instead of a summary for each CTR section. Given that the evaluation relies on using the same section with multiple statements, we need to generate only one summary per section if the summary does not depend on the hypothesis.

Inference. We solve the entailment task through zero-shot and two-shot prompting. As some statements might contain irrelevant sentences, we only keep the first sentence that contains the word "trial".

[4] razent/SciFive-large-Pubmed_PMC-MedNLI
[5] meta-llama/Llama-2-7b
[6] mistralai/Mistral-7B-Instruct-v0.2
[7] TheBloke/SOLAR-10.7B-Instruct-v1.0-GGUF

We are aware that it is not an ideal approach and we lose important information in some examples where 2 or more sentences are crucial for the task. We found that the model works better if we keep only the "trial" sentence rather than not performing this step at all. We also tried a more robust approach by asking the model to extract the relevant sentences from the hypothesis. While this tackles the issues encountered by our simple heuristic mentioned before, we limit our system setup to sentence splitting since it is significantly faster.

We limit the output grammar of the model to only "Yes" and "No" to ease processing. For each of the four sections and example types (Single or Comparison), we apply different prompts for summarization and evaluation. The final prompt templates are listed in Appendix B. The advantage of this approach is that we can analyze and tune the prompts independently for each section, without running the inference step for the whole dataset.

5 Results

Model	Setting	F1	Faithfulness	Consistency
PubMedBERT	fine-tuned	0.63	0.53	0.62
XLM-RoBERTa	fine-tuned	0.63	0.55	0.63
Deberta V3 Large	fine-tuned	0.63	0.54	0.62
BioBERT	fine-tuned	0.65	0.52	0.63
SciFive Pubmed PMC	zero-shot	0.56	0.61	0.56
Llama-2 7B	zero-shot	0.65	0.19	0.44
Mistral 7B	zero-shot	0.65	0.18	0.44
Mistral 7B Instruct-0.2	zero-shot	**0.72**	0.68	0.66
SOLAR 10B *	zero-shot	0.63	**0.90**	**0.72**
SOLAR 10B	few-shot	0.71	0.83	**0.72**

Table 1: Results of our submissions for the SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. Best results are presented in **bold**, and second-best results are presented in underline. Results with asterisk (*) were not submitted.

The official results of our team can be found in Table 1. Our results indicate that using pre-trained models on clinical text does not significantly improve the performance on this particular task, despite research confirming that domain-specific language model pre-training can be beneficial for other biomedical tasks (Gu et al., 2022). In line with last year's findings (Jullien et al., 2023b), we observe that instruction-tuned models pre-trained on generic datasets perform better than discriminative models pre-trained on biomedical datasets. With respect to LLaMa-2 and Mistral models used in zero-shot settings, they achieve high F1 scores of 0.65 and 0.72. However, these models are not robust enough and achieve low performance on Faithfulness and Consistency metrics, which are the metrics the organizers focused on. We further expand on the results of our best-performing model, SOLAR.

Control set. We obtain an F1-score of 0.71, with the highest score for comparisons of adverse events (0.79 F1) and the lowest score for comparisons of interventions (0.62 F1). Our team reaches the 16th place out of 32 participants in the official leaderboard. Compared to last year, we would be ranked on 5th place while using little to no training data and modest computational resources. Similar to last year, we report a higher Recall (0.73) than Precision (0.70).

Contrast set. Our system achieves a Faithfulness score of 0.83 (10th place out of 32 teams) and a Consistency score of 0.72 (14th place out of 32 teams). This shows that the model is more reliable when dealing with semantic-altering transformations compared to semantic-preserving changes, with only one CTR section having a faithfulness score below 0.79 (eligibility comparisons - 0.71). We observe a similar behavior in 22 other submissions where faithfulness is higher than consistency.

6 Error Analysis and Discussion

In this section, we analyze the types of errors made by the SOLAR model according to the provided metrics based on each intervention target.

Definition interventions. These interventions simply append a sentence to the statement. The model is capable of extracting the relevant sentence containing references to clinical trials if asked explicitly through a separate pre-processing step, but this incurs an additional runtime cost. Ultimately, we tackle this issue with a simple heuristic (see the inference details in 4.2).

Numerical interventions. Even though the SOLAR model has been tuned for mathematical reasoning, we identified several shortcomings. The model performs poorly with measurement units that express the same quantity in different ways. It appears to understand the meaning of symbols (e.g. "positive" instead of "+"), but if domain-specific acronyms are lowercase instead of uppercase (e.g. "hr", "her2", "mcs", "pdr"), the prediction changes.

Another interesting example is related to how entailment is affected when numbers have similar semantic meanings in other contexts. In one of the intervention trials, it is specified that a treatment

cycle takes 21 days. The statement asks whether the trial has a 30-day cycle, with the model classifying it as entailment. Slightly tweaking the number in the statement reveals that this is not regarded as entailment if the values do not match semantically: only the pairs 21-28, 21-30 and 21-31 are considered entailment; the other pairs from 21-26 to 21-34 are classified as contradiction. Nonetheless, if we change the number in the trial (21) instead of the statement, it always predicts an entailment, even when the value is nonsensical. This appears to be a drawback of long contexts as the issue only manifests with 2-shot prompting.

For numerical paraphrasing, we have 0.67 Consistency, 0.54 F1-score, 0.63 Recall and 0.47 Precision. Conversely, for numerical contradictions, the model obtains 0.80 Faithfulness, 0.85 Consistency, 0.91 F1, 1.0 Recall and 0.83 Precision[8].

Paraphrasing and contradiction interventions. We notice a similar behavior to the numerical interventions: very high scores for contradictions, but significantly lower results for rephrases. Thus, paraphrases have 0.70 Consistency, 0.69 F1, 0.72 Recall, and 0.66 Precision, while contradictions have 0.83 Faithfulness, 0.78 Consistency, 0.89 F1, 1.0 Recall, and 0.80 Precision.

Our results suggest the model is overly sensitive to any wording change between the hypothesis and premise, mistakenly interpreting them as conflicting. This also explains the perfect recall scores.

In the remainder of this section, we present a few high-level remarks related to our design decisions.

Few-shot prompting might improve results compared to zero-shot prompting. We used two examples from the training set, an entailment and a contradiction. This approach improved the results on average on the development set in terms of F1, regardless of model, model size or summarization configuration. Adding more examples to few-shot prompting might further increase the results at the expense of slower inference speed. However, with 6-shot prompting, the performance degrades even if we do not reach the context limit of 4096 tokens. We assume this is caused by the complexity of the task and the model "forgetting" what task it must solve. We do not further explore 6-shot prompting because of poor preliminary performance and increased runtime on a subset of the development data.

[8]We adapted the evaluation script to use `pos_label` as 0 for contradictions and numerical contradictions.

On the test set, it appears that few-shot prompting degrades the results in terms of Faithfulness and Contradiction, although it improves paraphrasing scores. We argue that a higher Faithfulness score in itself does not imply better results because the model could simply predict more Contradictions when the input is altered.

Instruction order is significant. We observed the best results when the instructions were placed at the start of the prompt, followed by the CTR (or CTR summary) and then the hypothesis. This strategy constantly provides improved predictions across most of the evaluated prompts. There is a drop of 4 F1 percentage points on the development set when the hypothesis is placed before the CTR summary in the prompt template. Separately, the predictions are also affected if the instruction is placed at the end of the prompt, after the hypothesis. Repeating the instruction before and after the CTR-hypothesis pair is as effective as simply placing the instruction before the CTR text.

Larger models obtain better results, but smaller models are still useful for prototyping. We use the 3-bit quantization version of the same SOLAR model (Q3_K_M) to experiment with more prompts, taking advantage of faster inference times. This approach has been very useful in designing the summarization prompts because the summarization step is the most expensive one in terms of computational resources. We also experimented with hybrid systems, where the summaries are generated with a smaller model and the inference task is done by a larger model. The performance of the hybrid strategy is comparable to running the entire pipeline with the larger model.

There are no hard constraints for summaries prompts, as long as they do not depend on the hypothesis. There appears to be no substantial difference in the final results when changing the prompt used to generate CTR summaries. We apply some of the following restrictions in each summarization prompt: use abbreviations, avoid verbs, use short sentences, be brief, maximum N words. Rephrasing the prompt does not seem to have a relevant impact for this task. As previously mentioned, it is paramount that the summaries preserve the meaning of the original text. The absence of relevant information in summaries is a major source of errors in our system. Unfortunately, for most sections the model was unable to create contextualized summaries conditioned by the hypothesis without mentioning the hypothesis in the summary.

For the Results section with a single CTR, conditioning the summary on the hypothesis looked promising. However, the next example from the test set confirms our concerns about this strategy, where a single summary for multiple hypotheses is not appropriate due to conditioning on the first hypothesis in the dataset. Our generated summary is: "There is no information in the given CTR report that relates to the statement about all patients treated with GTx-024 1mg gaining lean body mass over a 10 year period". The model is distracted by the "10 year period" from one of the hypotheses, altering the original meaning completely, even though the word "year" does not appear anywhere in the initial CTR section. See appendix A.1 for the full CTR text and associated hypotheses.

7 Conclusions and Future Work

In this paper, we presented our approach for the SemEval 2024 Task 2 (Jullien et al., 2024) aimed at understanding large language models behavior in clinical contexts. We explored several types of models and prompting techniques in order to determine whether fine-tuning is more feasible than zero-shot or few-shot prompting in a limited resource setting.

Our findings suggest that, while LLMs exhibit remarkable clinical NLI capabilities at a surface level, the proposed metrics and interventions uncover a tendency of the models to take shortcuts and rely on simple heuristics, especially when faced with semantic-preserving changes. We intend to investigate further methods of evaluating the reliability of large language models in future work.

To address the inherent weak numerical reasoning of our model (and all generative models), a promising strategy is to offload complex mathematical hypotheses to a specialized model like xVal (Golkar et al., 2023). This approach involves representing numbers as individual digits (e.g., 123 becomes ["1", "2", "3"]), replacing them with a generic [NUM] token, and scaling the token according to the original numerical value. Their results showed a 70-fold improvement over standard models. We can extend this strategy to create an ensemble where other weaknesses of our model (like rephrases) are offloaded to specialized models.

Acknowledgements

This work was partially supported by a grant on Machine Reading Comprehension from Accenture Labs and by the POCIDIF project in Action 1.2. "Romanian Hub for Artificial Intelligence".

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, et al. 2024. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. 2023. CHARD: Clinical health-aware reasoning across dimensions for text generation models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–327, Dubrovnik, Croatia. Association for Computational Linguistics.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. 2015. *Fundamentals of clinical trials*. Springer.

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. 2023. xVal: A continuous number encoding for large language models.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3:1–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *arXiv preprint arXiv:2310.04270*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling.

Dave Makhervaks, Plia Gillis, and Kira Radinsky. 2023. Clinical contradiction detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1248–1263, Singapore. Association for Computational Linguistics.

Bhavish Pahwa and Bhavika Pahwa. 2023. BpHigh at SemEval-2023 task 7: Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944, Toronto, Canada. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023. AutoTrial: Prompting language models for clinical trial design. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12461–12472, Singapore. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. MetaMath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

A Additional examples

A.1 Summary mismatch example

The Results section of the CTR with ID "NCT00467844" is the following:

> Outcome Measurement: The Efficacy of GTx-024 on Total Body Lean Mass. Change in total body lean mass as measured by dual energy x-ray absorptiometry (DEXA)from baseline to 4 months. Time frame: Baseline to Four Months.
>
> Results 1: Arm/Group Title: GTx-024 1 mg Arm/Group Description: [Not Specified] Overall Number of Participants Analyzed: 32 Median (Full Range) Unit of Measure: kg 1.55 (-2.06 to 12.64)
>
> Results 2: Arm/Group Title: GTx-024 3 mg Arm/Group Description: [Not Specified]

Overall Number of Participants Analyzed: 34 Median (Full Range) Unit of Measure: kg 0.98 (-4.84 to 11.54)

The summary is conditioned on the following hypothesis: "all patients treated with gtx-024 1mg in the primary trial gained lean body mass over a 10 year period". However, other hypotheses are concerned with other quantities: "at least one patient treated with GTx-024 1mg in the primary trial gained over 10 kilos of Lean body Mass". For the latter, our summary is misleading because the "10 kilos" information is missing. This could be mitigated by refraining to summarize short sections.

B Prompt templates

The prompt templates used to obtain the final leaderboard results for SOLAR are shown in tables 4 and 5.

We started with a single prompt template for all sections and summaries. When the results did not further improve, we analyzed the F1-score of each section, shown in Tables 2 and 3. Due to time constraints, we only create summaries for the first 50 examples in the train set.

While the single CTR summaries for the results section depend on the hypothesis, due to an implementation choice, all examples use the same CTR summary, regardless of the hypothesis (only the first hypothesis is used). We believe that this issue is not essential, since all the hypotheses for a CTR focus on the same information.

Initial summary prompt: "Instruction: You are given a clinical trial report. You must summarize the report. Use abbreviations. Be brief. Report: {premise}. Summary (max 350 words):".

Initial evaluation prompt: "Instruction: You are given a Clinical Trial Report and a hypothesis. ##Report: {premise}. ##Hypothesis: {hypothesis}. ##Can the hypothesis be inferred from the report? Respond only with Yes or No. ##Response (Yes or No):". This prompt was also used for our experiments with LLaMa-2 and Mistral.

C Infrastructure

In terms of infrastructure, we use a system with 16 GB RAM and an NVIDIA GTX 1060 MQ GPU with 6 GB VRAM. Out of the 5500 examples on the test set, this method only requires generating summaries for 251 examples comprising different CTR sections. On the development set, we need to create 100 summaries for the 200 samples, making the summarization step more expensive in this regard.

Section	F1
Eligibility (Single)	0.6637
Eligibility (Comparison)	0.5274
Intervention (Single)	0.7306
Intervention (Comparison)	0.7751
Results (Single)	0.7141
Results (Comparison)	0.7525
Adverse events (Single)	0.7141
Adverse events (Comparison)	0.6805

Table 2: Initial results for the train set (first 50 examples)

Section	F1
Eligibility (Single)	0.8178
Eligibility (Comparison)	0.8285
Intervention (Single)	0.7678
Intervention (Comparison)	0.6000
Results (Single)	0.7368
Results (Comparison)	0.7749
Adverse events (Single)	0.5835
Adverse events (Comparison)	0.6969

Table 3: Initial results for the development set

The inference time for a summary varies with the length of the CTR section, with a minimum time of 30 seconds per sample and a total time of about 7 hours, but it should be noted that this stage is a one-time cost. The inference time for one example is approximately 5 seconds, meaning that the evaluation on the test set takes roughly 9 hours, with the total time reaching 12 hours when applying 2-shot prompting, since the sequence length increases.

Section (CTR type)	Prompt
Eligibility (Single)	Instruction: You are given clinical trial criteria and a statement that may or may not be contradictory. Regarding the inclusion and exclusion criteria, is the statement correct? Respond only with Yes or No. ## Criteria: {premise}. ## Statement: {hypothesis}. ## Response (Yes or No):
Eligibility (Comparison)	Instruction: You are given clinical trial criteria for a primary and a secondary trial, and a statement. Regarding the inclusion and exclusion criteria, is the statement correct for each trial? Respond only with Yes or No. ## Criteria: {premise}. ## Statement: {hypothesis}. ## Response (Yes or No):
Intervention (Single)	Instruction: You are given a CTR and a statement. Can the statement be deduced from the CTR? Focus on the interventions. Respond only with Yes or No. ##CTR: {premise}. ##Statement: {hypothesis}. ##Response (Yes or No):
Intervention (Comparison)	(same prompt template as single CTR for interventions)
Results (Single)	Instruction: You are given the results of a CTR and a statement. Can the statement be deduced from the CTR in terms of number of participants, measures and results? Respond only with Yes or No. ##CTR: {premise}. ##Statement: {hypothesis}. ##Response (Yes or No):
Results (Comparison)	Instruction: You are given the results of a CTR and a statement. Can the statement be deduced from the CTR? Respond only with Yes or No. ##CTR: {premise}. ##Statement: {hypothesis}. ##Response (Yes or No):
Adverse events (Single)	Instruction: You are given a CTR and a statement. Regarding the adverse events, signs and symptoms observed in the CTR, is the statement correct? Respond only with Yes or No. ##CTR: {premise}. ##Statement: {hypothesis}. ##Response (Yes or No):
Adverse events (Comparison)	(same prompt template as single CTR for adverse events)

Table 4: Evaluation templates for each CTR section

Section (CTR type)	Prompt
Eligibility (Single)	Instruction: You are given the eligibility criteria for a clinical trial report. You must summarize the report focusing on inclusion and exclusion criteria. Report: {premise}. Use short sentences. Summary:
Eligibility (Comparison)	(same prompt template as single CTR for eligibility)
Intervention (Single)	Instruction: You are given the intervention information for a clinical trial report. Each report contains 1-2 cohorts, which receive different treatments, or have different characteristics. You must summarize the report focusing on the type, dosage, frequency, and duration of treatments being studied. Report: {premise}. Use short sentences to group by cohort. Summary:
Intervention (Comparison)	Instruction: You are given the intervention information for a clinical trial report. Each report contains 1-2 cohorts, which receive different treatments, or have different characteristics. You must summarize the report focusing on the type, dosage, frequency, and duration of treatments being studied. Report: {premise}. Use short sentences to group by cohort and group by primary trial and secondary trial. Summary:
Results (Single)	Instruction: You are given the results of a CTR and a statement. Extract all the relevant information from the CTR that is related to the statement. Report: {premise}. Statement: {hypothesis}. Answer:
Results (Comparison)	Instruction: You are given the results of two clinical trials. Each trial contains 1-2 cohorts, which receive different treatments, or have different characteristics. You must summarize the report for each trial, focusing on number of participants, outcome measures, units, results. Report: {premise}. Use short sentences and keep all numeric values. Summary:
Adverse events (Single)	Instruction: You are given the adverse events of a clinical trial report. Each report contains 1-2 cohorts, which receive different treatments, or have different characteristics. You must summarize the report focusing on the adverse events, signs and symptoms observed in patients. Report: {premise}. Use short sentences. Summary:
Adverse events (Comparison)	Instruction: You are given the adverse events of a clinical trial report. Each report contains 1-2 cohorts, which receive different treatments, or have different characteristics. You must summarize the report focusing on the adverse events, signs and symptoms observed in patients. Report: {premise}. Use short sentences and group by primary trial and secondary trial. Summary:

Table 5: Summarization templates for each CTR section. The results section (single CTR) is the only one for which summaries depend on the hypothesis due to lack of time to rerun the summaries for the test set.