IITK at SemEval-2024 Task 2: Exploring the Capabilities of LLMs for Safe Biomedical Natural Language Inference for Clinical Trials

Shreyasi Mandal **Ashutosh Modi**
Indian Institute of Technology, Kanpur (IIT Kanpur)
{shreyansi, ashutoshm}@cse.iitk.ac.in

Abstract

Large Language models (LLMs) have demonstrated state-of-the-art performance in various natural language processing (NLP) tasks across multiple domains, yet they are prone to shortcut learning and factual inconsistencies. This research investigates LLMs' robustness, consistency, and faithful reasoning when performing Natural Language Inference (NLI) on breast cancer Clinical Trial Reports (CTRs) in the context of SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. We examine the reasoning capabilities of LLMs and their adeptness at logical problem-solving. A comparative analysis is conducted on pre-trained language models (PLMs), GPT-3.5, and Gemini Pro under zero-shot settings using Retrieval-Augmented Generation (RAG) framework, integrating various reasoning chains. The evaluation yields an F1 score of **0.69**, consistency of **0.71**, and a faithfulness score of **0.90** on the test dataset.

1 Introduction

Clinical trials serve as essential endeavors to evaluate the effectiveness and safety of new medical treatments, playing a pivotal role in advancing experimental medicine. Clinical Trial Reports (CTRs) detail the methodologies and outcomes of these trials, serving as vital resources for healthcare professionals in designing and prescribing treatments. However, the sheer volume of CTRs (e.g., exceeding 400,000 and proliferating) presents a challenge for comprehensive literature assessment when developing treatments (Bastian et al., 2010). Natural Language Inference (NLI) (Bowman et al., 2015) emerges as a promising avenue for large-scale interpretation and retrieval of medical evidence bridging recent findings to facilitate personalized care (DeYoung et al., 2020; Sutton et al., 2020). The SemEval 2024 Task 2 on the Natural Language Inference for Clinical Trials (NLI4CT) (Jullien et al., 2024) revolves around annotating statements extracted from breast cancer CTRs[1] and determining the inference relation between these statements and corresponding sections of the CTRs, such as Eligibility criteria, Intervention, Results, and Adverse events. By systematically intervening in the statements, targeting numerical, vocabulary, syntax, and semantic reasoning, the task aims to investigate Large Language Models (LLM)s' consistency and faithful reasoning capabilities.

In this paper, we experiment with Gemini Pro (Team et al., 2023), GPT-3.5 (Brown et al., 2020), Flan-T5 (Longpre et al., 2023) and several pre-trained language models (PLMs) trained on biomedical datasets, namely BioLinkBERT (Yasunaga et al., 2022), SciBERT (Beltagy et al., 2019), ClinicalBERT (Huang et al., 2019). We conducted zero-shot evaluations of Gemini Pro and GPT-3.5, employing Retrieval Augmented Generation (RAG) framework (Lewis et al., 2020) integrating Tree of Thoughts (ToT) reasoning (Yao et al., 2023) facilitating multiple reasoning paths. Our experiments involved applying various instruction templates to guide the generation process. These templates were refined through manual comparison of the labels within the training dataset against those generated by the models. The PLMs were fine-tuned on the provided training dataset, while the Flan-T5 model was assessed under zero-shot conditions.

Gemini Pro emerged as the top-performing model among all the experimented models, achieving an F1 score of **0.69**, with consistency and faithfulness scores of **0.71** and **0.90**, respectively, on the official test dataset. Notably, a comparative analysis between GPT-3.5 and Gemini Pro revealed shortcomings in GPT-3.5's performance, particularly in instances requiring numerical reasoning. For detailed examination of such instances, please

[1] https://clinicaltrials.gov/ct2/home

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1397–1404

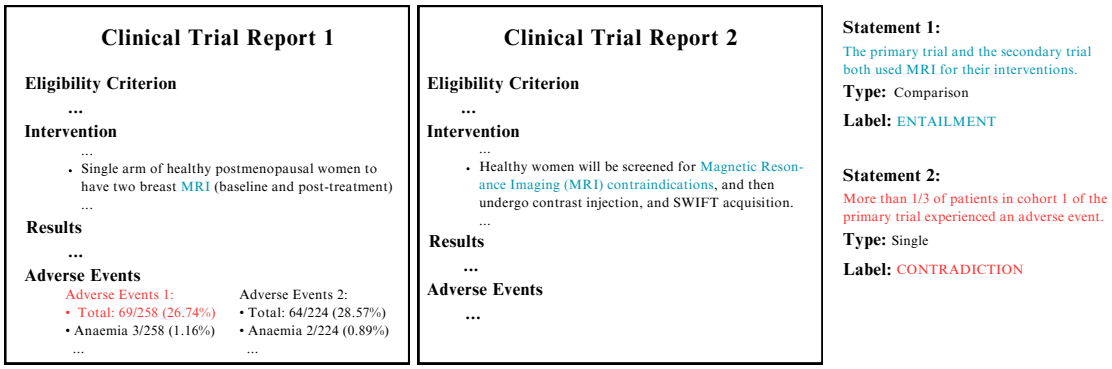

Figure 1: Examples of the dataset used in the NLI4CT task. Statement 1 compares the *Intervention* section from two different clinical trial reports, while statement 2 is based on the *Adverse Events* section of the first clinical trial report. The evaluation of the first statement requires textual inference skills, while the second requires numerical inference skills.

refer to Appendix A, where an example showcases GPT-3.5's accurate inference yet inadequate conclusion. The code to reproduce the experiments mentioned in this paper is publicly available.[2]

2 Background

2.1 Related Work

Pretrained Language Models (PLMs) and Large Language Models (LLMs) exhibit the potential to yield promising outcomes in the biomedical domain due to their ability to comprehend and process complex medical data effectively. BioLinkBERT (Yasunaga et al., 2022), pre-trained on PubMed[3], utilizes hyperlinks within documents. It has attained state-of-the-art (SOTA) performance across a wide range of tasks and various medical NLP benchmarks, namely BLURB (Gu et al., 2021) and BioASQ (Nentidis et al., 2020). SciBERT (Beltagy et al., 2019) is trained on scientific publications from the biomedical domain in Semantic Scholar[4]. ClinicalBERT (Huang et al., 2019) is trained using clinical text data sourced from approximately 2 million clinical notes contained within the MIMIC-III database (Johnson et al., 2016). Kanakarajan and Sankarasubbu (2023) employed a fine-tuned Flan-T5-xxl model with instruction tuning, achieving an F1 score of 0.834 on the SemEval 2023 Task 7 (Jullien et al., 2023a,b). Zhou et al. (2023) performed joint semantics encoding of the clinical statements followed by multi-granularity inference through sentence-level and token-level encoding, getting an F1 score of 0.856. Although these models have achieved high performance, there remains a need for further investigation into their application in vital areas such as real-world clinical trials.

GPT-3.5, developed by OpenAI[5] and comprising 175 billion parameters, uses alternating dense and locally banded sparse attention patterns in the transformer layers (Child et al., 2019; Wolf et al., 2020). The token size limit for GPT-3.5 (free tier) is 4,096. **Gemini Pro**, developed by Google DeepMind[6] uses decoder-only transformers (Vaswani et al., 2017) and multi-query attention (Shazeer, 2019) with a context window length of 32,768 tokens.

Data	Number of Samples	Labels	
		Entailment	Contradiction
train	1700	850	850
dev	200	100	100
test	5500	1841	3659

Table 1: The number of samples in each subset of the data. The distribution of the labels between the train and the development set is even. Note: The test set labels were made public after the completion of the task.

2.2 Task and Dataset Description

The NLI4CT task (Jullien et al., 2024) focuses on textual entailment based on a collection of breast cancer CTRs, statements, explanations and labels annotated by domain expert annotators. The CTRs are in English. The CTRs are segmented into four

[2]https://github.com/Exploration-Lab/IITK-SemEval-2024-Task-2-Clinical-NLI
[3]https://pubmed.ncbi.nlm.nih.gov
[4]https://www.semanticscholar.org
[5]https://openai.com
[6]https://deepmind.google

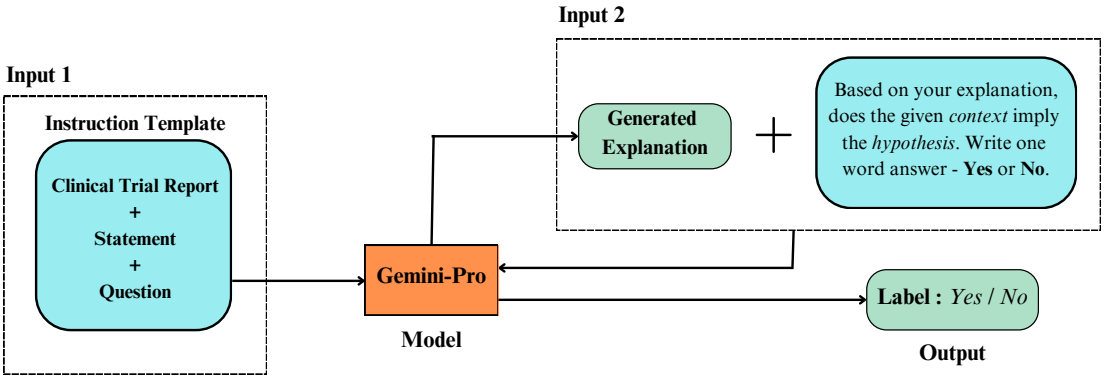

Figure 2: An overview of the proposed system architecture used for the NLI4CT Task

sections - eligibility criteria, intervention details, results, and adverse events. The statements, with an average length of 19.5 tokens, make claims about the information contained in one of the sections of a CTR or compare the same section from two different CTRs as seen in Figure 1. The task involves determining the inference relation (*entailment* or *contradiction*) between CTR-statement pairs. The dataset consists of 999 Clinical Trial Reports (CTRs) and 7400 annotated statements, which are divided into train, development and test sets. Table 1 provides statistics for the dataset.

3 System Overview

LLMs such as GPT-3 (Brown et al., 2020) and Gemini Pro (Team et al., 2023) have shown remarkable performances across various tasks. For the NLI4CT task, we have experimented with Gemini Pro, GPT-3.5, Flan-T5 (Longpre et al., 2023), BioLinkBERT (Yasunaga et al., 2022), SciBERT (Beltagy et al., 2019), ClinicalBERT (Huang et al., 2019) and ClinicalTrialBioBert-NLI4CT[7]. The performance of the different models is shown in Figure 7. Zero-shot evaluation was done on Gemini Pro and GPT-3.5, Flan-T5 was instruction fine-tuned following Kanakarajan and Sankarasubbu (2023), and the rest of the models were trained on the given train and development dataset. Gemini Pro and GPT-3.5 were considered for further experimentation because of their superior performance.

The proposed system utilizes structured instruction templates and multi-turn conversation techniques to generate explanations and labels for the statements provided as input, as shown in Figure 2.

Reasoning is an essential ability required by an LLM to solve complex problems (Qiao et al., 2022). Tree of Thoughts (ToT) framework (Yao et al., 2023) and Chain-of-Thought (CoT) reasoning (Wei et al., 2022) is integrated into the models, facilitating multiple reasoning paths.

3.1 Reasoning Frameworks

Chain-of-Thought (CoT) prompting (Wei et al., 2022) has demonstrated promising results in improving the reasoning abilities of LLMs. To evaluate Gemini Pro and GPT-3.5, we used Zero-shot-CoT (Kojima et al., 2022) prompt reasoning without requiring few-shot demonstrations. The phrase *"Let's think step by step"* is added after the instruction as shown in Figure 3.

```
"{premise}

{statement}
Is the given statement correct based on the information
provided above? Let's think step by step."

>> {explanation by model}

"From your answer above, is the statement true or false?
Write one word answer."

>> {True / False}
```

Figure 3: Instruction template for CoT prompting

Tree-of-Thought (ToT) framework (Yao et al., 2023; Long, 2023) relies on trial and error method to solve complex reasoning tasks. It facilitates multi-round conversations and backtracking. Our system allows for three reasoning paths using the prompt shown in Figure 4.[8]

For the evaluation of the model, the input to Gemini Pro and GPT-3.5 is constructed using an instruction template containing the appropriate prompt for ToT or CoT reasoning, data from the

[7] https://huggingface.co/domenicrosati/ClinicalTrialBioBert-NLI4CT

[8] https://github.com/dave1010/tree-of-thought-prompting

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.

Figure 4: Prompt for Tree of Thought reasoning

CTR which constitutes the premise and the statement or the hypothesis as shown in Figure 2. A series of two questions is presented to the model to generate both the explanation and the corresponding label. Multi-turn conversation (Zhang et al., 2018) is used to include the generated explanation as context for generating the final label. The explanation is also retained for further experimentations. The generated final label is converted as follows: *{"Yes": "Entailment", "No": "Contradiction"}*. A comparison of the performance of GPT-3.5 and Gemini Pro after integrating CoT and ToT reasoning frameworks is shown in Figure 6.

4 Experimental setup

4.1 Data Preprocessing

As discussed in Section 2.2, the statements can make claims about the information contained in one of the sections of a CTR, which is then called a "Single" statement or compare the same section from two different CTRs, called a "Comparison" statement. In "Single" statements, the term "primary" is employed to assert a claim. Evidence from the CTR is compiled into a unified text structure, exemplified as follows: *"For the primary trial participants, {primary evidences}"*. In contrast, for "Comparison" statements, the term "secondary" accompanies "primary". The evidences are then compiled as: *"For the primary trial participants, {primary evidences}. For the secondary trial participants, {secondary evidences}"*.

4.2 Hyperparameter Tuning

For Gemini Pro, the temperature of the model is set to 0.7 and the safety settings are set to "BLOCK_NONE". For GPT-3.5, the models *"gpt-3.5-turbo-0613"* and *"gpt-3.5-turbo-1106"* are used for experimentation among which *"gpt-3.5-turbo-0613"* performs considerably better. The temperature of the model is set to 0.6.

4.3 Prompt Engineering

The system was experimented with several prompts to improve its performance. The explanations generated by the model were examined manually to identify instances where the solution deviated from the correct path. The prompt *"You are a clinical expert and can seamlessly perform natural language inference"* was introduced to give the model an identity. Additionally, rules were enforced to confine the model's output within the provided context and to prevent hallucinations, achieved through the prompt: *"Please align with the context given and do not make any false assumptions of your own."* Furthermore, to integrate CoT reasoning within the ToT framework, the prompt *"Provide a step-by-step explanation of your thought process"* was introduced. The final instruction template is shown in Figure 5.

Several experiments were conducted to assess the model's performance on extracting the labels *"Entailment"* or *"Contradiction"* in the second question of the multi-turn conversation. The F1 scores for various prompts on the development set are presented in Table 2. Ultimately, Prompt 4 demonstrated the best performance and was chosen for the final pipeline.

Prompt	**F1 score**
Based on the comprehensive evaluation of the model's responses, is the given hypothesis deemed to be true or false? Write one word answer.	0.689
Does this imply that the given hypothesis is supporting the report or not? Give one word answer (Yes / No).	0.667
From your answer above, is the statement true or false? Write one word answer.	0.656
Based on your explanation, does the given context imply the hypothesis. Write one word answer.	**0.723**

Table 2: Performance of the model on the *dev* data for different prompts for extracting the labels

4.4 Evaluation Metrics

The NLI4CT task (Jullien et al., 2024) is evaluated on the basis of three metrics - F1 score, consistency and faithfulness. Faithfulness measures the accuracy of the system's predictions by evaluating its ability to predict outcomes for altered inputs correctly. If the model correctly adjusts its predictions

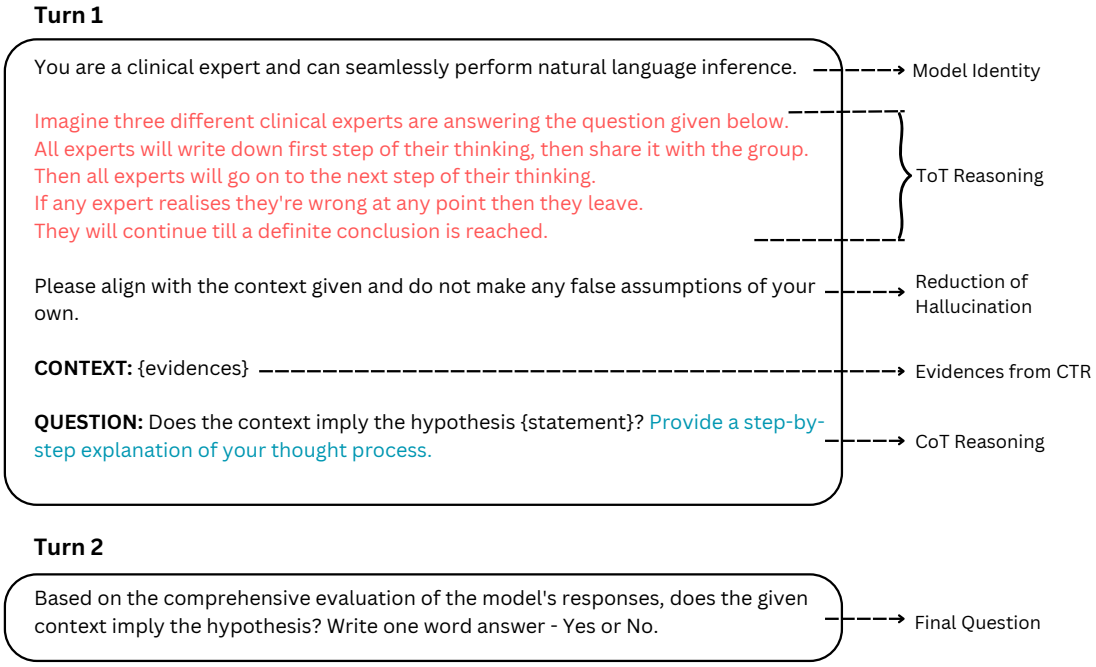

Figure 5: Final Instruction Template

in response to semantic alterations, it demonstrates higher faithfulness. On the other hand, consistency evaluates the model's ability to provide consistent predictions for semantically equivalent inputs.

5 Results

The zero-shot evaluation of Gemini Pro yielded an F1 score of **0.69**, with a consistency of **0.71** and a faithfulness score of **0.90** on the official test dataset. Our system achieved a fifth-place ranking based on the faithfulness score, a sixteenth-place ranking based on the consistency score, and a twenty-first-place ranking based on the F1 score. Gemini Pro outperforms GPT-3.5 with an improvement in F1 score by +1.9%, while maintaining almost similar consistency score. Additionally, the faithfulness score of Gemini Pro improves by +3.5% compared to GPT-3.5, as illustrated in Table 3.

Model	Base F1	Consistency	Faithfulness
Gemini Pro	**0.691**	0.712	**0.901**
GPT-3.5	0.672	**0.713**	0.866

Table 3: Results on the *test* data using Gemini Pro and GPT-3.5

The system utilizing Gemini Pro attained an F1 score of 0.72, while GPT-3.5 achieved an F1 score of 0.68 on the training dataset. Manual examination of the model-generated explanations and a comparison of the generated labels with the original labels was conducted to refine the prompts and enhance the model's responses.

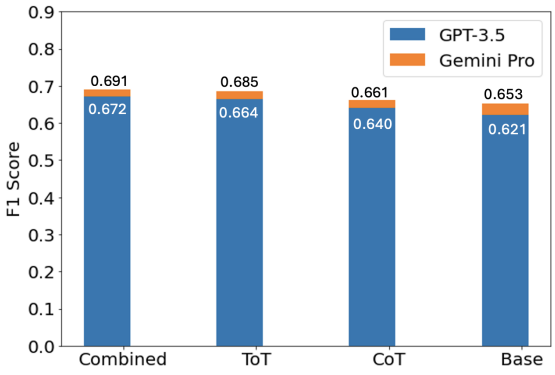

Figure 6: Comparison of the performance of Gemini Pro and GPT-3.5 without the integration of any reasoning framework, with CoT reasoning, with ToT reasoning and with both the reasoning frameworks combined.

As depicted in Figure 6, the integration of CoT reasoning led to an increase in performance for Gemini Pro and GPT-3.5 by 0.8% and 1.9%, respectively. Furthermore, upon integrating the ToT reasoning framework, the performance improved by 3.2% and 4.3%, respectively. When both ToT and CoT reasoning were integrated, the models showed an increase in performance by **3.8%** and **5.1%**, respectively, compared to the baseline model.

Figure 7 compares the performance of Gemini Pro and GPT-3.5, both without reasoning frameworks, with Flan-T5 and other experimented PLMs. Gemini Pro achieved the highest F1 score of 0.65, followed closely by GPT-3.5 with an F1 score of 0.62. Flan-T5 performed moderately with an F1 score of 0.57, while BioLinkBERT, SciBERT, ClinicalBERT, and CTBioBERT displayed lower F1 scores ranging from 0.46 to 0.53.

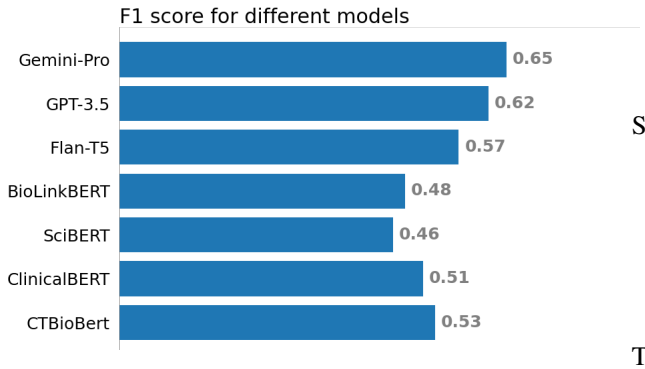

Figure 7: Performance (F1 Score) of the different experimented models. Note: CTBioBert represents the model ClinicalTrialBioBert-NLI4CT.

A comparative analysis between GPT-3.5 and Gemini Pro highlighted GPT-3.5's shortcomings in tasks requiring logical reasoning. Appendix A presents the example responses for both the models. The appendix further analyzes potential reasoning errors made by GPT-3.5 and Gemini Pro.

6 Conclusion

This paper presents an evaluation of several pretrained language models (PLMs), and GPT-3.5, Gemini Pro, under zero-shot conditions. Our analysis focuses on assessing the reasoning capabilities of GPT-3.5 and Gemini Pro and their adeptness at logical problem-solving. In the NLI4CT task, we achieved an F1 score of 0.691, consistency of 0.71, and faithfulness of 0.90. Additionally, our findings underscore that prompt engineering is crucial for large language models (LLMs). We have made our instruction templates and code publicly available to facilitate reproducibility.

7 Acknowledgments

We would like to thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments and constructive criticism. Their feedback has greatly helped us to improve the clarity and rigor of this work.

References

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

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.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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.

Kamal Raj Kanakarajan and Malaikannan Sankarasubbu. 2023. Saama ai research at semeval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 553–568. Springer.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

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):17.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. Thifly research at semeval-2023 task 7: A multi-granularity system for ctr-based textual entailment and evidence retrieval. *arXiv preprint arXiv:2306.01245*.

Appendix

A Example Responses

As illustrated in Figure 8, GPT-3.5 exhibits limitations in numerical reasoning tasks. Although steps 1 and 2 of the example are executed correctly, GPT-3.5 starts hallucinating in step 3, leading to illogical reasoning and ultimately yielding an incorrect conclusion that the statement is incorrect. When the same input is given to Gemini Pro, it correctly concludes that the statement entails the CTR as shown in Figure 9.

Step 1: Assess the exclusion criteria for the primary trial.
- The exclusion criteria mention a Karnofsky Performance status <90% as an exclusion criterion.
- This means that patients with a Karnofsky status below 90% would not be eligible for the trial.

Step 2: Compare the Karnofsky status of 94% mentioned in the hypothesis with the eligibility criteria.
- According to the hypothesis, patients with a Karnofsky status of 94% would be eligible for the primary trial.

Step 3: However, based on the exclusion criteria, a Karnofsky Performance status < 90% is not eligible, regardless of the specific value (i.e. 94%).

Conclusion: *"Patients with a Karnofsky status of 94% are eligible for the primary trial"* is incorrect.

Figure 8: Example of an incorrect response by GPT-3.5

Step 1: Check the inclusion criteria for Karnofsky Performance status.
Step 2: The inclusion criteria state that Karnofsky Performance status must be >= 90%.

Conclusion: *Patients with a Karnofsky status of 94% are eligible for the primary trial* because the inclusion criteria state that Karnofsky Performance status must be >= 90%.

Figure 9: Response by Gemini Pro for the same statement

Figure 10 showcases an excellent example of Gemini Pro's mathematical reasoning.

Hypothesis: Neutropenia affected the majority of patients in cohort 1 of the primary trial.
Label: Contradiction

Explanation:
...
From the context, we know that the total number of patients in cohort 1 of the primary trial is 69, and the number of patients who experienced neutropenia in cohort 1 of the primary trial is 4. Therefore, the percentage of patients in cohort 1 of the primary trial who experienced neutropenia is **4/69 * 100 = 5.8%**.

Since 5.8% is not the majority, the hypothesis is incorrect.

Figure 10: An example response by Gemini Pro showcasing its mathematical reasoning ability.