FZI-WIM at SemEval-2024 Task 2: Self-Consistent CoT for Complex NLI in Biomedical Domain

Jin Liu[1, 2] and Steffen Thoma[1]

[1]FZI Research Center for Information Technology, Karlsruhe, Germany
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
{jin.liu, thoma}@fzi.de

Abstract

This paper describes the inference system of FZI-WIM at the SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. Our system utilizes the chain of thought (CoT) paradigm to tackle this complex reasoning problem and further improves the CoT performance with self-consistency. Instead of greedy decoding, we sample multiple reasoning chains with the same prompt and make the final verification with majority voting. The self-consistent CoT system achieves a baseline F1 score of 0.80 (1st), faithfulness score of 0.90 (3rd), and consistency score of 0.73 (12th). We release the code and data publicly[1].

1 Introduction

The Safe Biomedical Natural Language Inference for Clinical Trials (NLI4CT) task aims to investigate the consistency and faithfulness of natural language inference (NLI) models in clinical settings(Jullien et al., 2024). NLI is a typical natural language task requiring natural language reasoning (Yu et al., 2023). Fine-tuned BERT-based (Devlin et al., 2019) discriminative models have been widely applied to solve NLI problems (Liu et al., 2020; He et al., 2021). Studies show increasing reasoning capabilities of large language models (LLMs), both in proprietary (Brown et al., 2020; Achiam et al., 2023) and open-source LLMs (Touvron et al., 2023a,b; Jiang et al., 2024). However, problems of inconsistency and unfaithfulness still occur with LLMs(Golovneva et al., 2023; Turpin et al., 2023). Compared to other domains, medical applications have much higher standards regarding safety and trustfulness, so inconsistencies and unfaithfulness limit AI applications in the medical domain.

Chain of Thought (CoT) has been proposed to elicit the reasoning capabilities of LLMs (Wei et al., 2022). Based on the CoT, further concepts like Tree of Thought (Yao et al., 2023a), ReAct (Yao et al., 2023b), Self-Consistency (Wang et al., 2023), and so on have been proposed to improve the performance of CoT further. One common characteristic of the frameworks mentioned above is the explicit generation of reasoning chains. Since only verification labels are provided in the NLI4CT training dataset, we utilize GPT-4 to generate reasoning chains. With the distilled knowledge from GPT-4, we further instruction-tune an open-source LLM with low-rank adaption (LoRA) (Hu et al., 2022) for claim verification with CoT. Our system follows the self-consistency concept by generating multiple CoT reasoning chains and verifying with majority voting.

We summarize our major findings regarding this task as follows:

- LoRA instruction-tuning can bring domain-specific (biomedical) knowledge and reasoning capabilities to LLMs.
- Our instruction-tuned LLM tends to contradict the statement if the information is only contained in the statement, not in the premise, even if the information is factually correct.
- CoT reasoning gains significant performance improvement regarding faithfulness compared to the label-only prediction.
- Compared to the greedy CoT, self-consistent CoT with majority voting has a performance improvement of 1.31 (baseline F1), 0.75 (consistency score), and 0.69 (faithfulness score) percentage points. Performance improvement is limited for the binary classification problem.

2 Background

NLI4CT contains one text entailment task, namely infer the relationship between a premise and a state-

[1]https://github.com/jens5588/FZI-WIM-NLI4CT

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1269–1279
June 20-21, 2024

Primary clinical trial report:
Outcome Measurement: Overall Tumor Response (OR) OR was defined as the percentage of participants experiencing either a confirmed complete response (CR) or a confirmed partial response (PR) according to Response Evaluation Criteria in Solid Tumors (RECIST) criteria 1.0. CR is defined as the disappearance of all lesions (target and/or non-target). PR is defined as at least a 30% decrease in the sum of the longest dimensions (LD) of target lesions taking as a reference the baseline sum LD, with non-target lesions not increased or absent. Time frame: Start of treatment to disease progression or death or discontinuation from study or at least 28 days after last dose (up to Week 131) Results 1: Arm/Group Title: Lapatinib 1000 mg + Nab-Paclitaxel Arm/Group Description: Participants received Lapatinib 1000 milligram (mg) tablets orally daily 1 hour before or after a meal along with a Nab-paclitaxel infusion at a dose of 100 mg/m^2 intravenously over 30 minutes on Day 1, 8, and 15, in a 4-week cycle, for at least 6 cycles. Overall Number of Participants Analyzed: 60 Measure Type: Number Unit of Measure: percentage of participants 53

Original Statement: Over 1/2 patients in the primary trial treated with Lapatinib 1000 mg + Nab-Paclitaxel experienced either a confirmed complete response (CR) or a confirmed partial response (PR). **Entailment**
Modified Statement 1: in the primary clinical trial, it was recorded that over 50% of those treated with lapatinib 1000 mg and nab-paclitaxel displayed either a verified complete or partial response. **Entailment**
Modified Statement 2: over 66% patients in the primary trial treated with lapatinib 1000 g + nab-paclitaxel experienced either a confirmed complete response (cr) or a confirmed partial response (pr). **Contradiction**

Figure 1: A data example. With the same clinical report, semantic-preserving and semantic-altering interventions on the original statement are used to evaluate the consistency and faithfulness of the verification system.

ment as either entailment or contradiction (Jullien et al., 2024). The premises in the NLI4CT dataset are collected from publicly available breast cancer clinical trial reports (CTRs), split into four sections: eligibility criteria, intervention, results, and adverse event (Jullien et al., 2023). The statements are sentences making claims about the information in the CTR premise, either about a single CTR or a comparison between 2 CTRs. The numbers of training, validation, and test examples are 1700, 200, and 5500, respectively. In the test set, 500 examples are used as anchors. The other 5000 statements are created with interventions on these first 500 examples. Figure 1 shows an example with interventions on the statements. Using the same clinical report, two statements are modified, one semantics-preserving (modified statement 1) and the other semantics-altering (modified statement 2), based on the original statement. The purpose of the interventions is to investigate the consistency and faithfulness of the inference.

For the text entailment task in the first iteration of NLI4CT (SemEval-2023 Task 7), most systems have fine-tuned discriminative transformer-based models (Jullien et al., 2023). Kanakarajan and Sankarasubbu (2023) instruction-tuned Flan-T5 model and achieved the 2nd place for the text entailment task. However, the systems mentioned above only predicted the verification labels without the reasoning process. To achieve a trustworthy verification, our system not only predicts the label but also the reasoning chains. Since the training and validation datasets have only provided verification labels. We utilize GPT-4 to verify with reasoning chains for training and validation datasets. We further instruction-tune an open-source LLM with distilled reasoning chains from GPT-4. To address the inconsistency problem in reasoning chains, we sample multiple chains and then employ a majority voting approach to determine the final verification outcome.

3 System overview

Figure 2 shows the pipelines of data creation, model training, and inference of our system. In the following, we describe each part.

3.1 Knowledge Distillation

The NLI4CT data only contains verification labels without rationales. Following the concept of knowledge distillation (Hinton et al., 2015), we leverage GPT-4 to generate rationales with CoT for the training and validation datasets. We further extract the verification labels of GPT-4 based on the CoT reasoning and filter out the examples for which the extracted labels are different from the gold labels.

3.2 LoRA Instruction-tuning

Parameter efficient fine-tuning (PEFT) gains popularity as the sizes of LLMs increase (Houlsby et al., 2019; Pfeiffer et al., 2020; Li and Liang, 2021; Hu et al., 2022). Low-rank adaption (LoRA) (Hu et al., 2022) is a PEFT approach to fine-tune LLMs. Studies show over-parametrized models reside on a low intrinsic dimension (Aghajanyan et al., 2021; Li et al., 2018; Hu et al., 2022). The key assumption of LoRA is that the updates to the weights also have a low intrinsic rank during adaptation for downstream tasks. The parameter updates of a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ can be represented as $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$ and $r \ll \min(d, k)$ (Hu et al., 2022).

Based on the concept of LoRA, given prompt x and the target output $y = (y_1, ..., y_m)$, the loss can be formulated as:

$$L = \sum_{i=1}^{m} -\log(p_\theta(\hat{y}_i = y_i | x, y_1, ..., y_{i-1})), \quad (1)$$

where θ represents W_0, B, A and only B and A are trainable.

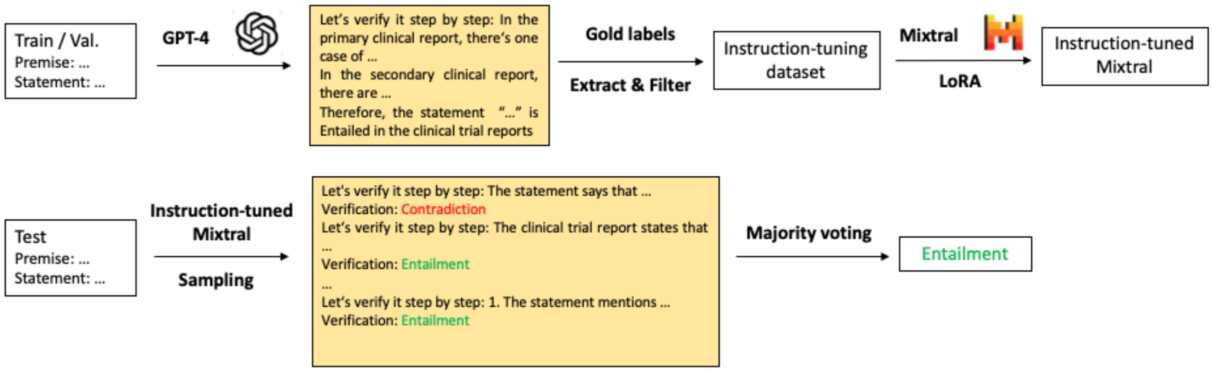

Figure 2: Training and Inference pipeline of self-consistent CoT system.

3.3 Self-Consistency

Self-consistency for LLMs was proposed by Wang et al. (2023) to replace the greedy decoding in CoT reasoning. The intuition behind self-consistency is that there are multiple ways to solve a complex problem. Another fact that supports the introduction of the self-consistency concept is that current LLMs still have difficulties with complex reasoning tasks and make mistakes in certain reasoning steps. With multiple reasoning chains, the model is less likely to make the same error and reach the same wrong answer. Wang et al. (2023) show self-consistency boosts the performance of different LLMs with different sampling strategies, including temperature sampling, top-k sampling, and nucleus sampling, on arithmetic, commonsense and symbolic reasoning tasks.

4 Experimental setup

In this section, we describe our experiment setup. We describe the CoT generation with GPT-4, the model training, and the inference setups.

4.1 LoRA Instruction-tuning

Instruction Data Creation Since GPT-4 has state-of-the-art reasoning capabilities, we instruct GPT-4 to verify the statements based on the CTR premise step-by-step for the training and validation datasets. The instruction prompt is shown in Appendix A.1. We then extract the verification labels of CoT rationales with an NLI model, namely bart-large-mnli (Lewis et al., 2020), and keyword matching. We compare the extracted labels with gold labels and filter out examples where the verification is wrong. For 1700 training and 200 validation examples, we achieve 1413 and 166 correctly verified examples with CoT, respectively, resulting in an accuracy of approximately 83%. With these selected examples, we build an instruction-tuning dataset for fine-tuning an open-source LLM. We show an example of CoT instruction-tuning data in Appendix A.2. For further comparison, we create another instruction dataset containing only verification labels without CoT for training and validation datasets. Appendix A.3 shows one label-only instruction-tuning data example.

Instruction-tuning We select Mixtral-8x7B-Instruct (Jiang et al., 2024) as our open-source LLM for instruction-tuning. The base model selection has considered reasoning capabilities and the number of model parameters. The configuration of LoRA is set as follows: r=8, alpha=32 and adaption of the attention weights of query (W_q), key (W_k), value (W_v) and output (W_o). This setup leads to 6.8M trainable parameters, which corresponds to 0.0146% of the base model size of 46.7B. We instruction-tune Mixtral-8x7-Instruct with the CoT instruction dataset and label-only instruction data separately for five epochs. Further implementation details are described in Appendix B.1.

4.2 Inference

Self-Consistent Inference To generate reasoning chains with the CoT instruction-tuned Mixtral model, we follow Wang et al. (2023) using a temperature sampling with the parameters $T = 0.7$ and $k = 50$ in top-k. For each (premise, statement) pair in the test dataset, we begin with sampling 10 reasoning chains. We then extract the verification labels and apply majority voting to decide the final label. If the result for both labels is tied, we further generate reasoning chains with different seeds. The maximal number of generated reasoning chains is 25. We show the number of generated reasoning chains in Appendix B.2.

Greedy Inference For comparison, we also apply the greedy decoding strategy to generate verification with label-only and CoT instruction-tuned models. With the greedy strategy, the models predict the next token with the highest probability without sampling[2].

4.3 Evaluation Metrics

Three evaluation metrics are used for the task: baseline F1, faithfulness score, and consistency score. The F1 score is used to evaluate the performance of the control set without intervention, consisting of 500 samples. There are 5000 samples in the test dataset, created with interventions based on the samples in the control set.

The consistency score evaluates whether the model predicts the same label when semantic-preserving interventions exist in the original statements (Jullien et al., 2024). For N statements x_i in the contrast set C, their corresponding original statements y_i and model prediction function f, the consistency score is calculated with:

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

where $x_i \in C$: Label$(x_i) =$ Label(y_i).

The faithfulness score evaluates whether the model can change its prediction when semantic-altering interventions are present in the original statements (Jullien et al., 2024). For N statements x_i in the contrast set C, their corresponding original statements y_i and model prediction function f, the faithfulness score is calculated with:

$$Faithfulness = \frac{1}{N}\sum_{1}^{N} |f(y_i) - f(x_i)| \quad (3)$$

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

5 Evaluation

In this section, we report the results of our self-consistent CoT system. For comparison, we also report the results of label-only verification and CoT-greedy. The comparison of three systems serves as an ablation study to analyze the performance improvement of extra steps, reasoning chain generation and self-consistent verification. Table 1 gives an overview of the three systems compared to the best scores in each category. Our self-consistent CoT system achieves 1st place regarding baseline F1, 3rd place regarding faithfulness, and 12th place regarding consistency. Next, we will analyze our system according to the three metrics separately.

[2] https://huggingface.co/docs/transformers/generation_strategies

Model	Base F1	Consistency	Faithfulness
Label-only Greedy	0.7867	**0.7364**	0.8102
CoT Greedy	0.7869	0.7217	0.8970
Self-Consistent CoT	**0.8000**	0.7292	**0.9039**
Best score	0.80	0.81	0.95

Table 1: Overview of three systems compared to the best scores of the task regarding each metric.

Model	Eligibility	Intervention	Adverse Events	Results
Label-only Greedy	0.7482	**0.8175**	0.8131	0.7679
CoT Greedy	0.7132	0.7794	**0.8667**	0.7961
Self-Consistent CoT	**0.7581**	0.7770	0.8305	**0.8485**

Table 2: F1 scores of three systems according to the sections in CTRs.

5.1 Baseline F1

In the test dataset, there are 500 samples without interventions on the statements, which are used as the control set. Table 2 shows the F1 scores of three systems according to the sections in the CTRs. The Eligibility section has the lowest F1 scores compared to the other sections in CTRs. Several factors have increased the difficulties for eligibility verification statements. First, the premises, i.e., criteria for inclusion and exclusion for clinical trials, are much longer than those for other sections. LLMs cannot always extract all relevant information from very long contexts. The verification of statements regarding eligibility often requires multi-step reasoning capabilities. The lack of domain-specific knowledge and common sense can also lead to verification errors. For the sections Intervention, Adverse Events, and Results, the major error type is numerical reasoning.

5.2 Consistency

Compared to the other two metrics, the consistency score of our self-consistent CoT system has the worst ranking. There are 4136 samples with semantic-preserving interventions, and they can be classified into five groups: paraphrase preserving, contradiction preserving, numerical paraphrase preserving, numerical contradiction preserving, and definitions preserving. Table 3 summarizes the consistency scores for each system according to

these categories. Compared to the other categories, self-consistent CoT and CoT greedy systems have worse performance regarding definitions and numerical paraphrase interventions. For the intervention regarding definitions, extra factual statements have been added to the original statements. A definition-intervention to the original statement in Figure 1, e.g., is:
Over 1/2 patients in the primary trial treated with Lapatinib 1000 mg + Nab-Paclitaxel experienced either a confirmed complete response (CR) or a confirmed partial response (PR). bladder solitary fibrous tumor is a solitary fibrous tumor that arises from the bladder. most tumors are benign.
Since the sentence *bladder solitary fibrous ...* is irrelevant to the clinical report, our self-consistent CoT tends to verify it as Contradiction (7 Contradictions : 3 Entailments in 10 generated reasoning chains).

For the category of numerical paraphrasing, there are 224 statements modified from 90 original statements. For the original set, we have around 81% accuracy (73 / 90), and for the modified set around 72% (162 / 224). We classify the errors, in total 34, where original statements are verified correctly and modified statements are verified wrong, into three groups: the conversion of fractional numbers, decimals, and percentages; the conversion of time, e.g., months to weeks, months to days, etc.; the conversion of units, e.g., mm to cm, mg to micrograms, etc. Among 34 misclassified statements, the system can generate at least one completely correct reasoning chain out of 10 reasoning chains for 17 statements. This indicates that the model has the knowledge for these conversions. However, our instruction-tuned model has difficulties utilizing this knowledge. The model also has difficulties understanding some paraphrased statements, e.g., from *over 1/2 patients* to *over 0.5 patients*, from *5% of patients* to *0.05 of patients*, etc. Another issue is some conversions are not quite precise, e.g., *9 months* to *36 weeks*, *6 months* to *180 days*. The model often contradicts the modified statements since they are not the same (9 months equals about 39 weeks).

5.3 Faithfulness

The contrast set for evaluating the faithfulness of the system is generated from 250 samples from the control set. All the 250 statements in the control set have the label Entailment. With contradiction intervention, 864 samples are created with

	Consistency				
Model	**para.**	**cont.**	**num. para.**	**num. cont.**	**defi.**
Label-only Greedy	0.7807	0.7613	**0.7411**	0.9568	**0.6553**
CoT Greedy	0.7780	0.8427	0.7321	0.9568	0.5780
Self-consistent CoT	**0.8020**	**0.8440**	0.7232	**0.9877**	0.5720

Table 3: Consistency scores of three systems according to semantic-preserving intervention types

	Faithfulness	
Model	**cont. alter.**	**num. cont. alter.**
Label-only Greedy	0.8040	0.8509
CoT Greedy	0.8933	0.9211
Self-consistent CoT	**0.8987**	**0.9386**

Table 4: Faithfulness scores of three systems according to intervention (altering) types

the label Contradiction, 114 of which are numerically intervened. Table 4 summarizes the faithfulness scores of each system after contradiction intervention types. The scores show that both CoT greedy and self-consistent CoT have a significant performance improvement over label-only prediction. This underscores the critical role of the extra reasoning chain generation step for faithful verification.

5.4 Self-consistent CoT and CoT Greedy

According to Table 1, our self-consistent CoT has improved the performance of the CoT greedy system regarding all metrics, namely baseline F1 of 1.31 percentage points, consistency score of 0.75 percentage points, and faithfulness score of 0.69 percentage points. However, the improvement is insignificant compared to Wang et al. (2023). We show two examples in Appendix C. The first example shows that self-consistent CoT corrects the error in CoT greedy, and the second one shows that self-consistent CoT fails to correct the error. One possible reason for the insignificant improvement is the number of generated reasoning chains. Due to computational limitations, we have only generated 10.36 reasoning chains on average, which is much less than the 40 reasoning chains in Wang et al. (2023). Another important reason is the aggregating mechanism of majority voting. From the above-mentioned second example, we can see the late simple aggregation of the verification labels is not enough for the binary classification problem. A more finegrained verification and integration of intermediate reasoning steps of CoT is needed to tackle the inconsistency problem further.

6 Conclusion

In this paper, we have reported our self-consistent CoT system for the SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. For comparison, we also reported the label-only greedy and CoT greedy inference systems. To achieve a trustworthy inference system, we utilized the CoT reasoning paradigm, not only predicting the verification labels but also rationales. We tackled CoT's inconsistency problem with self-consistent CoT. Compared to the greedy CoT system, we have improved the inference performance by generating multiple reasoning chains and verifying with majority voting. However, the performance improvement is limited. For future work, a more fine-grained evaluation of the correctness of the reasoning steps in the CoT paradigm is promising for solving complex reasoning tasks.

7 Acknowledgments

This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the project "DeFaktS" (Grant 16KIS1524K). This work was also supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. 2023. Gpt-4 technical report.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. 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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and et al. 2024. Mixtral of experts.

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.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *ArXiv*, abs/1804.08838.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural language reasoning, a survey.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.

A Prompts

A.1 Instruction prompt for GPT-4

Figure 3 shows an instruction example for GPT-4 to verify the statements in the train and validation datasets with CoT.

<s>[INST] **Primary clinical trial report**:
Adverse Events 1: Total: 12/32 (37.50%) Anaemia 0/32 (0.00%) Neutropenia 1/32 (3.13%) Thrombocytopenia 4/32 (12.50%) Atrial fibrillation 1/32 (3.13%) Cardiac failure congestive 1/32 (3.13%) Myocardial ischaemia 1/32 (3.13%) Abdominal discomfort 0/32 (0.00%) Ascites 1/32 (3.13%) Constipation 0/32 (0.00%) Rectal haemorrhage 1/32 (3.13%) Vomiting 1/32 (3.13%) Fatigue 1/32 (3.13%) Adverse Events 2: Total: 8/20 (40.00%) Anaemia 1/20 (5.00%) Neutropenia 0/20 (0.00%) Thrombocytopenia 1/20 (5.00%) Atrial fibrillation 0/20 (0.00%) Cardiac failure congestive 0/20 (0.00%) Myocardial ischaemia 0/20 (0.00%) Abdominal discomfort 1/20 (5.00%) Ascites 0/20 (0.00%) Constipation 2/20 (10.00%) Rectal haemorrhage 0/20 (0.00%) Vomiting 0/20 (0.00%) Fatigue 0/20 (0.00%)
Statement: Most of the cases of CHF in the primary trial, were in cohort 1.
Verify whether the statement is entailed in the primary clinical trial report with Entailment or Contradiction. [/INST]
Let's verify it step by step:

Figure 3: Example prompt for GPT-4

<s>[INST] **Primary clinical trial report**:
Adverse Events 1: Total: 12/32 (37.50%) Anaemia 0/32 (0.00%) Neutropenia 1/32 (3.13%) Thrombocytopenia 4/32 (12.50%) Atrial fibrillation 1/32 (3.13%) Cardiac failure congestive 1/32 (3.13%) Myocardial ischaemia 1/32 (3.13%) Abdominal discomfort 0/32 (0.00%) Ascites 1/32 (3.13%) Constipation 0/32 (0.00%) Rectal haemorrhage 1/32 (3.13%) Vomiting 1/32 (3.13%) Fatigue 1/32 (3.13%) Adverse Events 2: Total: 8/20 (40.00%) Anaemia 1/20 (5.00%) Neutropenia 0/20 (0.00%) Thrombocytopenia 1/20 (5.00%) Atrial fibrillation 0/20 (0.00%) Cardiac failure congestive 0/20 (0.00%) Myocardial ischaemia 0/20 (0.00%) Abdominal discomfort 1/20 (5.00%) Ascites 0/20 (0.00%) Constipation 2/20 (10.00%) Rectal haemorrhage 0/20 (0.00%) Vomiting 0/20 (0.00%) Fatigue 0/20 (0.00%)
Statement: Most of the cases of CHF in the primary trial, were in cohort 1.
Verify whether the statement is entailed in the primary clinical trial report with Entailment or Contradiction. [/INST]
Let's verify it step by step: The statement refers to cases of Cardiac Heart Failure (CHF) in the primary trial. Looking at the data from the primary clinical trial report, we can see that in Adverse Events 1 (cohort 1), there was 1 case of Cardiac failure congestive out of 32 participants, which is 3.13%. In Adverse Events 2 (cohort 2), there were 0 cases of Cardiac failure congestive out of 20 participants, which is 0.00%.
Therefore, the statement "Most of the cases of CHF in the primary trial, were in cohort 1" is Entailed in the primary clinical trial report.
Verification: Entailment </s>

Figure 4: An example of CoT instruction-tuning dataset

A.2 Example for CoT instruction-tuning dataset

Figure 4 shows an example of CoT instruction-tuning dataset. The rationales in CoT are generated by GPT-4 and we add the final verification label to CoT for convenience by extracting the labels.

A.3 Example for label-only instruction-tuning dataset

Figure 5 shows an example of the label-only instruction-tuning dataset, where only the gold label is appended to the (premise, statement) pair without CoT rationales.

B Implementation Details

B.1 Instruction-Tuning

For the instruction-tuning, we have applied PyTorch Fully Sharded Data Parallel (FSDP) from Meta AI (Zhao et al., 2023). The training script is adapted based on llama-recipes[3]. We train Mixtral-8x7B-Instruct on a cluster with 16 A100-40GB GPUs with a batch size of 1 and a learning rate of 0.0001. Mixed precision and gradient checkpointing have been applied to save memory. Evaluated on the validation dataset, the model achieves the best performance with 5 epochs. Following Kanakarajan and Sankarasubbu (2023), we include the validation set as part of the training data for the final submission.

B.2 Model Inference

We use the transformers[4] library to generate responses with instruction-tuned models. There are two setups for generation, greedy decoding and temperature sampling. Under temperature sampling, we set $T = 0.7$ and $k = 50$ in top-k. For each pair, we set `num_return_sequence` as 10. Occasionally, there are duplicates in the returned sequences. If the result is tied with majority voting, we further generate reasoning chains with different seeds. The maximum number of generated chains for one pair is 25. Table 5 shows the distribution for numbers of generated distinct reasoning chains in the test dataset. On average, we generate 10.36 reasoning chains for each pair in the test dataset.

[3] https://github.com/facebookresearch/llama-recipes/tree/main

[4] https://github.com/huggingface/transformers

<s>[INST] **Primary clinical trial report**:
Adverse Events 1: Total: 12/32 (37.50%) Anaemia 0/32 (0.00%) Neutropenia 1/32 (3.13%) Thrombocytopenia 4/32 (12.50%) Atrial fibrillation 1/32 (3.13%) Cardiac failure congestive 1/32 (3.13%) Myocardial ischaemia 1/32 (3.13%) Abdominal discomfort 0/32 (0.00%) Ascites 1/32 (3.13%) Constipation 0/32 (0.00%) Rectal haemorrhage 1/32 (3.13%) Vomiting 1/32 (3.13%) Fatigue 1/32 (3.13%) Adverse Events 2: Total: 8/20 (40.00%) Anaemia 1/20 (5.00%) Neutropenia 0/20 (0.00%) Thrombocytopenia 1/20 (5.00%) Atrial fibrillation 0/20 (0.00%) Cardiac failure congestive 0/20 (0.00%) Myocardial ischaemia 0/20 (0.00%) Abdominal discomfort 1/20 (5.00%) Ascites 0/20 (0.00%) Constipation 2/20 (10.00%) Rectal haemorrhage 0/20 (0.00%) Vomiting 0/20 (0.00%) Fatigue 0/20 (0.00%)
Statement: Most of the cases of CHF in the primary trial, were in cohort 1.
Verify whether the statement is entailed in the primary clinical trial report with Entailment or Contradiction. [/INST]
Verification: Entailment </s>

Figure 5: An example of label-only instruction-tuning dataset

Number of Chains	**Number of Samples**
10	5150
9	104
15	101
20	90
25	36
19	8
17	5
16	2
8, 14, 18, 24	1

Table 5: Distribution of numbers of generated distinct reasoning chains for test dataset

C Reasoning Chains Generation

Figure 6 shows an example where CoT greedy has wrong verification while self-consistent CoT has correct verification with majority voting. Figure 7 shows an example where both CoT greedy and self-consistent CoT have wrong verification.

<s>[INST] **Primary clinical trial report**: INTERVENTION 1: Arm 1 Herceptin/navelbine INTERVENTION 2: Arm 2 Taxotere/carboplatin/herceptin
Statement: only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention
Verify whether the statement is entailed in the primary clinical trial report with Entailment or Contradiction. [/INST]
Let's verify it step by step:

Gold Label: Contradiction

CoT Greedy:
The statement mentions that only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention. In the clinical trial report, Herceptin is mentioned in both Intervention 1 and Intervention 2. Herceptin is a brand name for trastuzumab, a monoclonal antibody used to treat certain types of breast cancer. Therefore, the statement is Entailed in the primary clinical trial report.
Verification: Entailment

Self-consistent CoT
Generation 1:
The statement mentions that only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention. Trastuzumab is a monoclonal antibody used to treat certain types of breast cancer. It is also known by its brand name, Herceptin. In the primary clinical trial report, Herceptin is mentioned in both Intervention 1 (Arm 1) and Intervention 2 (Arm 2). Therefore, both arms of the trial are using trastuzumab (Herceptin) in their interventions. So, the statement is a Contradiction to the information provided in the primary clinical trial report.
Verification: Contradiction
Generation 2:
The statement mentions that only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention. Looking at the clinical trial report, we can see that Herceptin is used in both arms of the trial. Herceptin is a brand name for trastuzumab, a medication used to treat certain types of breast cancer. Therefore, the statement is a Contradiction because both arms of the trial use trastuzumab, not just one.
Verification: Contradiction
Generation 3:
The statement mentions that only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention. Trastuzumab is the generic name for Herceptin. Looking at the clinical trial report, we see that Herceptin is indeed used in both interventions. In Intervention 1, Herceptin is used in combination with navelbine. In Intervention 2, Herceptin is used in combination with Taxotere and carboplatin. Therefore, the statement contradicts the information given in the primary clinical trial report. The correct answer is Contradiction.
Verification: Contradiction
Generation 4:
The statement mentions that only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention. In the clinical trial report, Herceptin is mentioned in both Intervention 1 and Intervention 2. Herceptin is a brand name for trastuzumab, a monoclonal antibody used to treat certain types of breast cancer. Therefore, the statement is Entailed in the primary clinical trial report.
Verification: Entailment
Generation 5:
The clinical trial report mentions two interventions. In Intervention 1 (Arm 1), Herceptin is mentioned. Herceptin is a brand name for the drug trastuzumab. In Intervention 2 (Arm 2), there is no mention of Herceptin or trastuzumab. Therefore, the statement "only one arm of the primary clinical trial makes use of trastuzumab as part of its intervention" is Entailed in the primary clinical trial report. Verification: Entailment
......

Figure 6: An example of self-consistent CoT successfully corrects the error in CoT greedy verification. We only show 5 generations of self-consistent CoT. The ratio of 10 self-consistent CoT generations is 8 Contradictions: 2 Entailments.

<s>[INST] **Primary clinical trial report**: Adverse Events 1: Total: 10/30 (33.33%) Hemoglobin decreased 2/30 (6.67%) Abdominal pain 1/30 (3.33%) Colitis 1/30 (3.33%) Diarrhea 7/30 (23.33%) Nausea 2/30 (6.67%) Rectal hemorrhage 1/30 (3.33%) Fatigue 1/30 (3.33%) Skin infection 1/30 (3.33%) Neutrophil count decreased 1/30 (3.33%) Platelet count decreased 3/30 (10.00%) Dehydration 1/30 (3.33%)
Secondary clinical trial report: Adverse Events 1: Total: 3/6 (50.00%) Anaemia 0/6 (0.00%) Febrile neutropenia 0/6 (0.00%) Neutropenia 0/6 (0.00%) Bradycardia 0/6 (0.00%) Diarrhoea 0/6 (0.00%) Pancreatitis 0/6 (0.00%) Vomiting 0/6 (0.00%) Disease progression 0/6 (0.00%) Fatigue 0/6 (0.00%) Pyrexia 0/6 (0.00%) Cholelithiasis 0/6 (0.00%) Hepatic pain 0/6 (0.00%) Bacteraemia 0/6 (0.00%) Adverse Events 2: Total: 3/6 (50.00%) Anaemia 0/6 (0.00%) Febrile neutropenia 0/6 (0.00%) Neutropenia 1/6 (16.67%) Bradycardia 0/6 (0.00%) Diarrhoea 0/6 (0.00%) Pancreatitis 0/6 (0.00%) Vomiting 0/6 (0.00%) Disease progression 0/6 (0.00%) Fatigue 1/6 (16.67%) Pyrexia 1/6 (16.67%) Cholelithiasis 0/6 (0.00%) Hepatic pain 1/6 (16.67%) Bacteraemia 0/6 (0.00%)
Statement: diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants
Verify whether the statement is entailed in the clinical trial reports with Entailment or Contradiction. [/INST]
Let's verify it step by step:

Gold Label: Contradiction

CoT Greedy:
In the primary clinical trial report, the occurrence of diarrhea is 7/30 (23.33%). In the secondary clinical trial report, the occurrence of diarrhea is 0/6 (0.00%) in both Adverse Events 1 and Adverse Events 2. Therefore, the statement "diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants" is Entailed in the clinical trial reports.
Verification: Entailment

Self-consistent CoT
Generation 1:
The statement claims that diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants. Looking at the primary clinical trial report, it shows that the incidence of diarrhoea was 7 out of 30 participants, which is approximately 23.33%. In the secondary clinical trial report, the incidence of diarrhoea is 0 out of 6 participants in both Adverse Events 1 and Adverse Events 2, which is approximately 0.00%. Comparing these two reports, it is clear that diarrhoea was indeed more common in the primary trial participants than in the secondary trial participants. Therefore, the statement is a Contradiction to the information provided in the clinical trial reports.
Verification: Contradiction
Generation 2:
The statement says that diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants. Looking at the primary clinical trial report, Diarrhea was reported in 7 out of 30 participants, which is approximately 23.33%. In the secondary clinical trial report, Diarrhoea was reported in 0 out of 6 participants in both Adverse Events 1 and Adverse Events 2, which is 0%. Comparing these two, it is clear that Diarrhoea was indeed more common in the primary trial participants than in the secondary trial participants. Therefore, the statement is a Contradiction to the information provided in the clinical trial reports.
Verification: Contradiction
Generation 3:
In the primary clinical trial report, the incidence of diarrhea was 7/30 (23.33%). In the secondary clinical trial report, the incidence of diarrhoea was 0/6 (0.00%) for both Adverse Events 1 and Adverse Events 2. Therefore, the statement "diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants" is Entailed in the clinical trial reports.
Verification: Entailment
Generation 4:
In the primary clinical trial report, the occurrence of diarrhea is reported as 7/30 (23.33%). In the secondary clinical trial report, the occurrence of diarrhea is reported as 0/6 (0.00%) in both Adverse Events 1 and Adverse Events 2. Therefore, the statement "diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants" is Entailed in the clinical trial reports.
Verification: Entailment
Generation 5:
In the primary clinical trial report, Diarrhea is reported in 7 out of 30 participants, which is approximately 23.33%. In the secondary clinical trial report, Diarrhoea is reported in 0 out of 6 participants in both Adverse Events 1 and Adverse Events 2, which is 0.00%. Therefore, the statement "diarrhoea was uncommon among the primary trial participants compared to the secondary trial participants" is Entailed in the clinical trial reports.
Verification: Entailment
......

Figure 7: An example of self-consistent CoT fails to correct the error in CoT greedy verification. We only show 5 generations of self-consistent CoT. The ratio of 10 self-consistent CoT generations is 7 Entailments: 3 Contradictions.