Advancing Biomedical Claim Verification by Using Large Language Models with Better Structured Prompting Strategies

Siting Liang[1], **Daniel Sonntag**[1,*]
[1]German Research Center for Artificial Intelligence, Germany
[*]University of Oldenburg, Germany
siting.liang|daniel.sonntag@dfki.de

Abstract

Biomedical claim verification involves determining the entailment relationship between a claim and evidence derived from medical studies or clinical trial reports (CTRs). In this work, we propose a structured four-step prompting strategy that explicitly guides large language models (LLMs) through (1) claim comprehension, (2) evidence analysis, (3) intermediate conclusion, and (4) entailment decision-making to improve the accuracy of biomedical claim verification. This strategy leverages compositional and human-like reasoning to enhance logical consistency and factual grounding to reduce reliance on memorizing few-shot exemplars and help LLMs generalize reasoning patterns across different biomedical claim verification tasks. Through extensive evaluation on biomedical NLI benchmarks, we analyze the individual contributions of each reasoning step. Our findings demonstrate that comprehension, evidence analysis, and intermediate conclusion each play distinct yet complementary roles. Systematic prompting and carefully designed step-wise instructions not only unlock the latent cognitive abilities of LLMs but also enhance interpretability by making it easier to trace errors and understand the model's reasoning process. This research aims to improve the reliability of AI-driven biomedical claim verification.

1 Introduction

Natural language inference (NLI) tasks typically involve determining whether or not a given hypothesis is entailed with respect to a premise (Bowman et al., 2015). An NLI system labels the logical relationship between the premise and hypothesis (e.g., Entailment, Contradiction, or Neutral). To enhance transparency and trustworthiness, the system should also provide an explanation in the form of specific evidence (rationales) that justify its decision (Camburu et al., 2018). In the scientific and medical domains, NLI is used to assist clinicians and researchers by automatically verifying claims against evidence from clinical trial data or medical literature. Specifically, it requires a deep understanding of medical and scientific knowledge to interpret implicit data points beyond simple text matching.

Clinical trial data often contain complex statistical information and precise measurements that must be interpreted accurately to avoid errors in claim verification. One example from the **NLI4CT** challenges (Jullien et al., 2023) shown in Figure 1 highlights the significant difficulties of applying NLI to validate statements (hypotheses) related to clinical trial reports (CTRs), which requires more than simple textual analysis. To accurately assess the claims, NLI models must process long and complex documents while also comprehending domain-specific terminology and applying multi-hop reasoning to draw connections that are not immediately obvious (Romanov and Shivade, 2018; Wadden et al., 2020; Jullien et al., 2024).

Large language models (LLMs) offer promising potential to address these challenges. Recent research has shown that the reasoning capability of LLMs depends on two key factors: the size of the model and the appropriateness of the prompts provided for specific tasks (Huang and Chang, 2022; Qiao et al., 2022). Using structured, multi-step prompting methods has been the subject of research efforts to explore the reasoning abilities of LLMs in different tasks, including mathematical problems, commonsense reasoning and multi-hop question answering tasks (Wei et al., 2022; Zhou et al., 2022; Xia et al., 2024). Larger LLMs excel in zero-shot reasoning but require careful prompt engineering for reliability and interpretability (Kojima et al., 2022; Jeblick et al., 2024). Smaller models offer faster inference with lower computational costs and are more suitable for real-time applications, though they have weaker reasoning abilities and rely more

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 148–166
August 1, 2025 ©2025 Association for Computational Linguistics

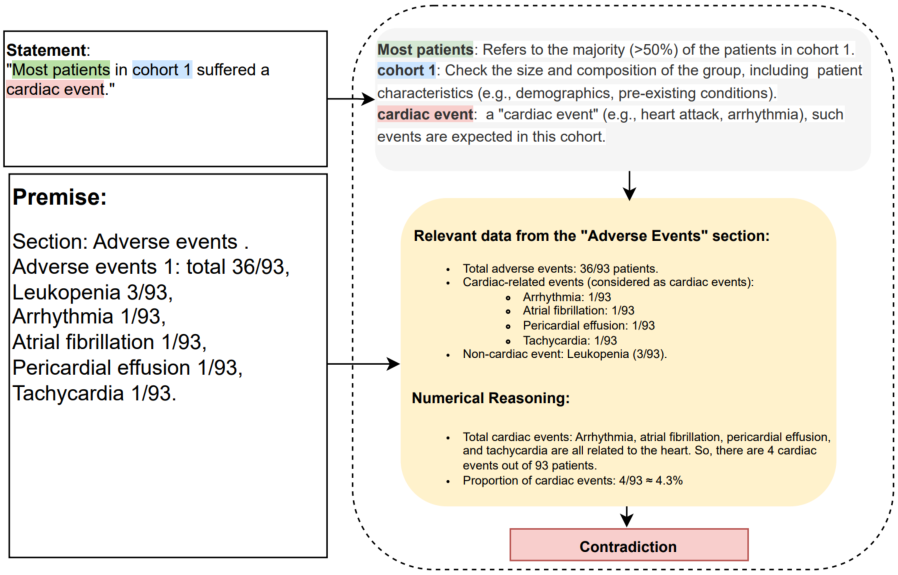

Statement:
"Most patients in cohort 1 suffered a cardiac event."
Premise:
Section: Adverse events .
Adverse events 1: total 36/93,
Leukopenia 3/93,
Arrhythmia 1/93,
Atrial fibrillation 1/93,
Pericardial effusion 1/93,
Tachycardia 1/93.
Most patients: Refers to the majority (>50%) of the patients in cohort 1.
cohort 1: Check the size and composition of the group, including patient characteristics (e.g., demographics, pre-existing conditions).
cardiac event: a "cardiac event" (e.g., heart attack, arrhythmia), such events are expected in this cohort.
Relevant data from the "Adverse Events" section:
Total adverse events: 36/93 patients.
Cardiac-related events (considered as cardiac events):
Arrhythmia: 1/93
Atrial fibrillation: 1/93
Pericardial effusion: 1/93
Tachycardia: 1/93
Non-cardiac event: Leukopenia (3/93).
Numerical Reasoning:
Total cardiac events: Arrhythmia, atrial fibrillation, pericardial effusion, and tachycardia are all related to the heart. So, there are 4 cardiac events out of 93 patients.
Proportion of cardiac events: 4/93 ≈ 4.3%
Contradiction

Figure 1: An example from **NLI4CT** dataset (Jullien et al., 2023). Left: a pair of statement and the adverse event section of a clinical trial data (premise). Right: an illustration of understanding the key terms and reasoning capabilities required to infer the logical relationship between the statement and the premise.

on fine-tuning for domain-specific tasks (Abdin et al., 2024).

In this work, we evaluate a 4-step NLI-oriented prompting framework that systematically decomposes the biomedical claim verification process into sequential stages, enhancing the zero-shot reasoning capabilities of LLMs, as illustrated in Figure 3. In particular, we aim to address the challenges posed by the need for domain expertise and the extensive length of medical documents, as well as the demand for the reliability. In our experiments, we specifically investigate the improvement obtained in both lightweight LLMs (from 3.6 to 14 billion parameters at most) and GPT3.5 and GPT-4o-mini models (OpenAI, 2024). Notably, all investigated models demonstrate a significant improvement over the standard CoT method, achieving an approximate 10% performance gain. Our key contributions can be summarized into:

- We investigate a 4-step prompting framework that extends the zero-shot CoT methodology and aims to address complex reasoning tasks like biomedical claim verification.

- We demonstrate how combining our structured prompting framework and supervised fine-tuning (SFT) significantly enhances the performances of lightweight LLMs in tackling the biomedical claim verification tasks.

The code for reproducing our experiments is available in the GitHub repository .

2 Approach

Task Definition We frame the NLI task in the biomedical domain as a binary classification problem, where an NLI system based on LLMs determines whether a statement or claim (C) logically follows from the premise (P) provided in clinical trial or scientific study data. For automatic performance evaluation, the final output of the system is a prediction of the logical relationship between C and P. Let's denote:

$$f(C, P) = \begin{cases} \text{Entailment} & \text{if } C \text{ logically follows} \\ & \text{ from } P; \\ \text{Contradiction} & \text{otherwise} \end{cases} \tag{1}$$

The binary prediction accuracy provides a straightforward measure of the LLMs' reasoning capabilities.

For solving the task, our first baseline utilizes a straightforward prompt template (see Fig 2), as proposed by Jullien et al. (2024). For clarity, in this paper, we will use "claim" and "statement" interchangeably to refer to the hypothesis within the context of NLI for biomedical claim verification, as different benchmarks employ varying conventions for these terms.

Zero-shot 2-step CoT Intermediate steps are useful for increasing grounded context and intermediate steps also increase the reliability of model generations (Yu et al., 2023). Standard CoT is a prompting methodology guiding LLMs to handle reasoning tasks by mimicking the thoughts of

Prompt template

```
"Given a section of 2 clinical
trial descriptions and a statement,
determine whether the statement
logically follows from the sections.
If the statement logically follows
from the sections, you need to
return 'Entailment'.  If the
statement does not logically follow
from the sections, you need to
return 'Contradiction'. The output
should be a single word <Entailment>
or <Contradiction>.
"Statement: " + Statement
"Primary Trial: " + Primary CTR text
"Secondary Trial: " + Secondary CTR
text
```

Figure 2: A simple prompt template for **NLI4CT** task.

solving example tasks demonstrated in prompts (Brown, 2020; Wei et al., 2022). While breaking down complex reasoning tasks into simpler steps can be useful, Zhou et al. (2022) noted that decomposition prompts require task-specific design for optimal performance.

In biomedical claim verification cases, providing multiple human-annotated examples in prompts is impractical due to the length of input documents, which can individually exceed 5,000 tokens. Furthermore, adding examples along with model-generated responses for intermediate steps would exceed the model's input limits and introduce noise to harm performance. As an alternative baseline, we adopt the zero-shot CoT approach (Kojima et al., 2022), which we refer to as the 2-Step prompting strategy in our experiments. In the first step, the model is prompted with an instruction phrase "step by step" instead of examples to generate a CoT response that leads to a solution. In the second step, the response from the first step is used to prompt the model to produce an output. Based on the task-specific prompt template as shown in Fig 2), our zero-shot CoT baseline follows a 2-Step prompting framework, utilizing instructive prompts in the first step: *'Determine whether the statement logically follows from the sections step by step.'*. The prompt text in step 2 includes the response generated in the first step followed by the remaining part from the template shown in Fig 2, e.g. *'If the statement ...The output should be a single word <Entailment> or <Contradiction>.'*.

4-step prompting framework In biomedical claim verification, each claim requires a distinct focus—some may involve analyzing a trial's inclusion criteria, while others require verifying the adverse event count in the outcomes section. While the 2-Step prompting method can be effective in simpler contexts, we identify key limitations of this approach when generalized to biomedical claim verification, particularly when using lightweight LLMs.

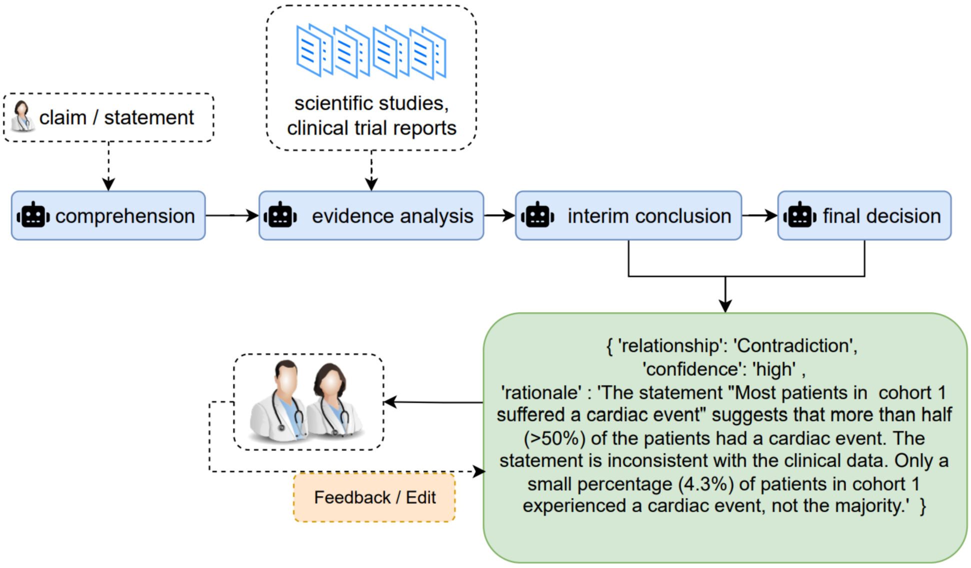
- Lack of co-reference resolution of terms or abbreviations between statement and premise data, leading to misinterpretation of key terms in the reasoning process.

- Resulting in shallow analysis without addressing each relevant factual detail in the premise (see some example generations of different models in Appendix A).

To address the challenges posed by complex biomedical terminology and diverse reasoning patterns, we draw from prior research in context-aware reasoning and domain-specific inference to develop a structured and adaptable prompting approach. We propose a carefully designed CoT framework with four sequential steps to improve vague reasoning of LLMs (see Fig 3). Below, we explain the intention of each step in more detail.

- **Claim Comprehension**: In the first step, the model only receives the statement and a targeted prompt instructing the model to interpret the medical terminology and complex biomedical concepts within the statement, e.g. *"Interpret the key terms in the statement based on biomedical knowledge. "*. This step serves to activate relevant domain knowledge and establish a semantic context for associating relevant information in later stages.

- **Evidence Analysis**: After understanding the statement, the model is presented with the premise data in the second step, such as original text from a clinical trial or medical study. The model is instructed to identify the relevant data points as evidence from the source compared to the information in the statement. Thus, the model focuses on verifying the truth of the statement by identifying the relevant evidence and performing comparative analysis. This analysis may involve numerical reasoning or biomedical reasoning, depending on the understanding of semantic context of each instance in the previous and current



{ 'relationship': 'Contradiction',
'confidence': 'high' ,
'rationale' : 'The statement "Most patients in cohort 1 suffered a cardiac event" suggests that more than half (>50%) of the patients had a cardiac event. The statement is inconsistent with the clinical data. Only a small percentage (4.3%) of patients in cohort 1 experienced a cardiac event, not the majority.' }

Figure 3: The proposed structured reasoning framework for with LLMs biomedical claim verification. In each step, we have the dedicate instruction for the model to complete the

steps. Example instruction at this stage include: *"1. Identify the relevant data points. 2. Evaluate each piece of information in the statement against these data points."*. The response generated in this stage serves as the basic for logical deduction in the subsequent inference stages, reducing issues like premature judgment.

- **Interim Conclusion**: LLMs likely draw conclusions in the evaluation response. However, these conclusions often lead to diverse outputs if lack task-specific focus. Therefore, the conclusion step builds on the tendency of LLMs to generate conclusions in their response but explicitly guides the models to focus on deducing logical relationship. For instance, we provide the following prompt in the third step: *"Conclude the evidence and determine whether the statement logically follows from the clinical trial data."* . This instruction refines the conclusion of the evaluation and steers the model response to explicitly determining the logical relationship.

- **Entailment Decision-making**: The final step encapsulates the model's reasoning path in a single relation prediction in natural words, e.g. *"Entailment"* or *"Contradiction"* as it is shown in the prompt template (Fig 2). This relationship prediction provides a concise outcome, enabling effective evaluation with automated metrics calculation.

By structuring the biomedical claim verification task into well-defined steps and emphasizing semantic grounding and evidence-based evaluation before logical inference, our approach helps LLMs focus on specific subtasks, reducing ambiguity and enhancing accuracy.

3 Experiments

Our experiments aim to address the main research question:

- How effectively does the 4-step strategy enhance the performance of LLMs in complex numerical and domain-specific reasoning tasks, particularly in biomedical claim verification?

Datasets Our primary evaluation task in this work is **NLI4CT** (Jullien et al., 2024), which presents challenges in numerical and domain-specific knowledge reasoning, as illustrated in Fig 1. Additionally, we assess the generalization capabilities using two related benchmarks: **SciFact** (Wadden et al., 2020) and **HealthVer** (Sarrouti et al., 2021). Both **SciFact** and **HealthVer** were designed as NLI tasks. While the claims in **SciFact** are written by human experts given scientific study abstracts of focusing medical research, the claims of **HealthVer** are directly extracted from studies. The premises in both datasets consist of evidence sentences extracted from relevant studies, requiring models to assign a relation label—*Support* or *Refute*—between input claims and the sentence-level premises. Wadden et al. (2022) highlighted the limitations of relying solely on sentence-level premises for scientific claim verification and demonstrated the advantages of incorporating document-level premises. For our experiments, we use the versions of **SciFact** and **HealthVer** provided by Wadden et al. (2022), which link each claim-premise pair to its relevant study source. To align with our task definition, we exclude the negative samples where the studies lack sufficient information to determine whether the claims are *Entailed* or *Contradicted*. Furthermore, we omit experiments involving the **CovidFact** (Saakyan et al., 2021) dataset due to the issues with noisy claims, including ungrammatical statements or claims unrelated to the provided sources (Wadden et al., 2022). Table 1 summarizes the instance distribution for each relation class applied in our evaluation.

Dataset	**Entailment/Support**	**Contradiction**
NLI4CT(test set)	250	250
SciFact (dev set)	216	122
HealthVer (test set)	503	308

Table 1: Number of instances in three different datasets for zero-shot experiments. **SciFact**'s test set withholds ground truth labels for leaderboard submissions, here we use its dev set as substitute.

Metrics For evaluating the performance of LLMs in our task, we employ the F1-score as the key evaluation metric for binary classification results.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Models While GPT-4 demonstrates advanced reasoning capabilities on the NLI4CT task in a zero-shot setting without specific prompting strategies (Gema et al., 2024), its closed-source nature and high cost make it impractical for experimenting with different prompting methods. Given computational constraints, our experiments prioritize small-scale, cost-effective LLMs that maintain competitive performance. We employ instruction-tuned (Ouyang et al., 2022) lightweight open-source LLMs (Abdin et al., 2024; Jiang et al., 2023; Team et al., 2024; Dubey et al., 2024) that are compatible with the *FastLanguageModel* Modules of unsloth.ai (Unsloth, 2024) for faster running and fine-tuning with LoRA method (Hu et al., 2021) on a single *NVIDIA* A100–80GB GPU. Table 7 provides the version information about the models utilized in our experiments, including comparisons with two low-cost GPT models: GPT-4o-mini and GPT3.5 (OpenAI, 2024).

Data Augmentation for Supervised Fine-Tuning While the proposed prompting strategy can enhance the performance of LLMs in logical inference, a significant performance gap still exists between larger and smaller LLMs. The limitations of smaller models include difficulties in producing responses with the correct format and challenges in controlling response length (Ding et al., 2023). To fine-tune small-scale LLMs, high-quality training examples are essential. The zero-shot performance of the GPT-4o-mini model demonstrates its potentials to generate such data without human-written inference examples (Gilardi et al., 2023). The second research question in our experiments is:

- Can fine-tuning improve the reliability and consistency of the output of the small-scale LLMs using GPT-4o-mini generated samples within the 4-step prompting framework?

We employ GPT-4o-mini to generates examples using the NLI4CT train set. If the model's final output deviates from the human-annotated label from the train set, e.g. predicting a *Contradiction* when the correct label is *Entailment*, the model is prompted to refine its reasoning in the second step to reach the correct logical conclusion. The prediction results are presented in Table 2). The 4-step responses generated by GPT-4o-mini achieve 0.98 accuracy on the entailment classification task using 1,700 samples from the NLI4CT train set. These high-quality responses can be confidently used to fine-tune small-scale LLMs.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Entailment	0.99	0.97	0.98	850
Contradiction	0.97	0.99	0.98	850
Accuracy			0.98	1700
Macro Average	0.98	0.98	0.98	1700
Weighted Average	0.98	0.98	0.98	1700

Table 2: Entailment classification performance of GPT-4o-mini on the NLI4CT train set (1,700 samples) using a 4-step prompting strategy with an additional refinement step verifying against the ground truth labels.

4 Results

4.1 Zero-Shot Results

Our results in Table 3 highlight the substantial performance gains achieved by both cost-effective

Model	**NLI4CT**			**SciFact**			**HealthVer**		
	Simple	2 Steps	**4 Steps**	Simple	2 Steps	**4 Steps**	Simple	2 Steps	**4 Steps**
GPT3.5	0.52 ± 0.01	0.53 ± 0.00	0.75 ± 0.01	0.51 ± 0.03	0.76 ± 0.00	0.86 ± 0.00	0.51 ± 0.01	0.60 ± 0.01	0.74 ± 0.02
GPT-4o-mini	0.67 ± 0.01	0.77 ± 0.02	0.86 ± 0.01	0.83 ± 0.01	0.88 ± 0.00	0.94 ± 0.01	0.69 ± 0.02	0.72 ± 0.01	0.77 ± 0.02
Phi3.5-3.6B	0.53 ± 0.00	0.61 ± 0.01	0.66 ± 0.02	0.51 ± 0.01	0.70 ± 0.03	0.80 ± 0.02	0.51 ± 0.01	0.70 ± 0.01	0.72 ± 0.01
Mistral-7B	0.55 ± 0.01	0.59 ± 0.02	0.69 ± 0.00	0.50 ± 0.02	0.72 ± 0.02	0.80 ± 0.02	0.44 ± 0.02	0.70 ± 0.00	0.72 ± 0.02
Llama3.1-8B	0.47 ± 0.00	0.54 ± 0.01	0.67 ± 0.02	0.53 ± 0.02	0.80 ± 0.01	0.84 ± 0.05	0.44 ± 0.02	0.70 ± 0.00	0.72 ± 0.01
Gemma2-9B	0.63 ± 0.00	0.67 ± 0.03	0.75 ± 0.03	0.57 ± 0.01	0.73 ± 0.00	0.86 ± 0.02	0.65 ± 0.02	0.70 ± 0.02	0.74 ± 0.01
Mistral-12B	0.55 ± 0.00	0.65 ± 0.01	0.75 ± 0.01	0.65 ± 0.01	0.83 ± 0.00	0.87 ± 0.02	0.50 ± 0.02	0.72 ± 0.00	0.74 ± 0.01
Phi3-14B	0.62 ± 0.01	0.64 ± 0.00	0.75 ± 0.02	0.76 ± 0.03	0.80 ± 0.01	0.88 ± 0.02	0.68 ± 0.02	0.72 ± 0.01	0.75 ± 0.01

Table 3: F1 Scores (mean ± standard deviation) for three benchmarks in zero-shot scenario. We compare the performance across the cost-effective GPT models and open sourced lightweight LLMs.

commercial models and small-scale LLMs when utilizing the 4-step prompting framework. Compared to the simple prompt template and 2-step baselines, the 4-step approach enhances reasoning quality and classification accuracy, demonstrating its effectiveness in zero-shot entailment tasks.

Ablation The four Steps starts with *claim comprehension*, where the model interprets the main claim and key terms in the statement. Without this initial step, the comparative analysis process at the *evidence analysis* stage, which involves *"identifying relevant data points and evaluating the information in the statement against these data points"*, likely results in reasoning paths that are less coherent. The ablation results in Fig 4 demonstrate that the absence of this comprehension step can hinder the accuracy of LLMs in claim verification tasks.

Model	**SciFact**		**HealthVer**	
	zero-shot	SFT*	zero-shot	SFT*
Phi3.5-3.6B	0.80	0.85	0.72	0.74
Mistral-7B	0.80	0.87	0.72	0.74
Llama3.1-8B	0.84	0.89	0.72	0.74
Gemma2-9B	0.86	0.90	0.74	0.75
Mistral-12B	0.87	0.88	0.74	0.75
Phi3-14B	0.88	0.90	0.75	0.77

Table 4: A comparison of F1 Scores (mean) for related tasks in zero-shot scenario and SFT(SFT* only with NLI4CT training samples).

4.2 Supervised Fine-Tuning Results

Fig 5 shows that supervised fine-tuning (SFT) with a small number of examples significantly improves F1-scores for lightweight LLMs, with performance further increasing as the number of training instances grows. Notably, Llama3.1-8B exhibits the largest performance gains, benefiting the most from the fine-tuning process.

Table 4 presents the generalization performance of lightweight models fine-tuned with **NLI4CT** samples, evaluated on the related tasks.

We observe that SFT significantly advantages the quality of *evidence analysis* in the second step, which is the primary contributor to the improved results. See some example responses of the small-scale models in zero-shot setting and after SFT in Appendix from table 10 to 15. Moreover, SFT improves task-specific control by ensuring adherence to specific instructions and maintaining a consistent response format, such as JSON, thereby enhancing the LLM's reliability not only for in-domain task - **NLI4CT**, but also the related tasks: **SciFact** and **HealthVer**. These improvements highlight the effectiveness of integrating structured reasoning with clear instructions of subtasks for enhancing smaller models in complex reasoning tasks like biomedical claim verification.

5 Discussion

Incorporation of GPT-4o-mini Our prompting approach underscores the importance of evidential evaluation in the second step in biomedical claim verification tasks. As shown in Fig 5, fine-tuning lightweight LLMs with step 2 responses generated by GPT-4o-mini significantly improves their performance on the **NLI4CT** task. Similarly, Table 5 demonstrates the positive impact of incorporating GPT-4o-mini's responses during the *evidence analysis* stage within the 4-step framework. These findings indicate that leveraging GPT-4o-mini's robust reasoning capabilities enhances the evidential evaluation process, enabling smaller LLMs to generate more accurate outputs. Whether to fine-tune lightweight LLMs with GPT-4o-mini generated data or to integrate GPT-4o-mini's evaluations directly into the 4-step pipeline depends on the specific requirements, computational constraints, and operational objectives of the application.

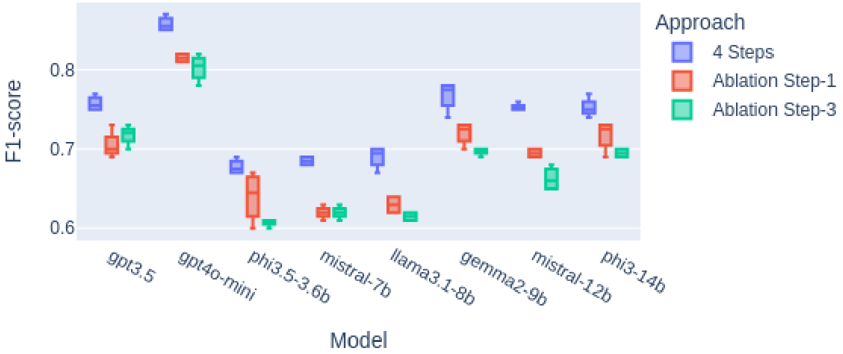


Figure 4: Omitting Step 1 or Step 3, LLMs result in degraded performance.

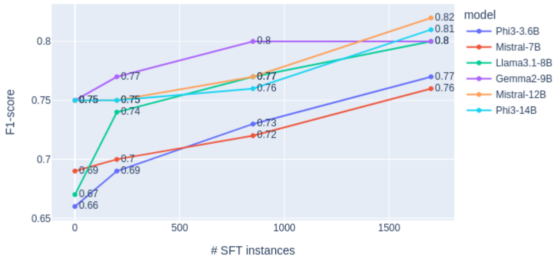


Figure 5: F1-scores of various lightweight LLM models with increasing numbers of SFT instances from **NLI4CT** train set.

	GPT-4o-mini in 2.Stage	**SFT***
Phi3-3.6B	0.80	0.77
Mistral-7B	0.84	0.76
Llama3.1-8B	0.84	0.80
Gemma2-9B	0.84	0.80
Mistral-12B	0.84	0.82
Phi3-14B	0.85	0.82

Table 5: **SFT*** are the results of each lightweight models fine-tuned within the 4-step framework using GPT-4o-mini generated training samples.

Error Analysis Our 4-step prompting framework enables the diagnosis of different types of error in LLMs by tracing the source of errors in individual steps.

- In the **Evidence Analysis** step: There are general mistakes the models fall into: numerical reasoning errors and misinterpretation of clinical trial data. Numerical reasoning errors involve misreading or miscalculating quantitative information. As it is shown in example 12 , interpreting *"98% had recurrence-free survival"* as the opposite, or failing to recognize that 98% of 100 patients implies that only 2 did not meet the outcome. Misinterpretation of clinical trial data occurs when the model overemphasizes irrelevant details or introduces unwarranted assumptions, such as requiring subgroup information when the data clearly applies to the overall population. These errors reflect the model's overconfidence in its generated explanations and highlight the difficulty large language models face when performing factual, logic-based reasoning.

Logical Conclusion
While it is true that the placebo group had a higher incidence of emesis compared to the Aprepitant group, the claim that the placebo group was "twice as likely" to experience emesis is misleading. The actual ratio of emesis incidence between the two groups is approximately 2.14 times (75% divided by 35%), which does not equate to a simple doubling. Therefore, the statement is not fully accurate as it oversimplifies the comparative risk and does not reflect the nuanced findings of the trial data. **Relation Prediction:** Contradiction
The data indicates that patients receiving the placebo intervention were approximately 2.14 times more likely to experience emesis compared to those receiving Aprepitant. Therefore, the statement that "Patients receiving the placebo intervention in the primary trial were twice as likely to experience emesis as patients receiving Aprepitant" can be inferred from the provided data, confirming its validity. **Relation Prediction:** Entailment

Table 6: Divergent conclusions and corresponding logical relation predictions.

- In the **Interim Conclusion** step: as shown in Table 6, we observed two opposite logical conclusions: *contradiction* and *entailment*, produced by different LLMs for the same verification sample illustrated in Fig 3. Despite being instructed and exposed to the same evaluation results, these models arrived at divergent conclusions. This highlights the necessity of further preference exploration to better align model outputs with user intent (Ouyang et al., 2022; Rafailov et al., 2024).

Pilot User Study As presented in Figure 3, a feedback-driven loop would support the development of collaborative systems that balance the responsibility for decision making between users and LLMs. Since collecting real-world feedback from clinicians and medical experts is challenging, we conducted a pilot user study with four medical students to assess the transparency and accountability of the model's reasoning process leading to the final relation classification on 20 examples from **NLI4CT** test set. Participants rated the model-generated responses on a scale from 1 (*'reasoning process was confusing'*) to 5 (*'the reasoning was easy to follow and correct'*). All participants rated the model responses as a **4**, suggesting that the model's reasoning process is generally perceived as transparent and confident. When asked what could be improved, participants provided feedback indicating the need for better quality in the intermediate reasoning steps generated by the LLMs, i.e. *"The model sometimes overlooked the smallest details in the claim."*. This highlights how enhanced interpretability can help identify limitations in reasoning of LLM. Also, as emphasized by (Huang et al., 2024), improving the functionality of these model-generated explanations is crucial for fostering user confidence in the system.

6 Related Work

Chain-of-thought Reasoning in LLMs Leveraging massive amounts of training data and billions of parameters, LLMs have demonstrated enhanced performance in various reasoning tasks. In particular, Chain-of-Thought (CoT) strategies (Wei et al., 2022), which provide exemplars of clear, step-by-step reasoning processes have demonstrated impressive performance in guiding LLMs to complete various reasoning tasks. Kojima et al. (2022) further showed that zero-shot CoT prompting, using the simple instruction LET'S THINK STEP BY STEP. instead of explicit examples, can also elicit strong reasoning capabilities from LLMs. However, their performance can vary depending on the complexity of the task and form of reasoning (Huang and Chang, 2023). Lei et al. (2023) addresses ungrounded misinformation in language model outputs by checking for factual inconsistencies between model generation and source documents at the sentence and entity levels within a chain of NLI framework. Zhou et al. (2022) involves breaking down complex problems into a series of simpler sub-problems, with the final problem being addressed depending on the responses to earlier sub-problems, and has proved generalization across different tasks. The evolution of CoT and CoX methodologies (Zhou et al., 2022; Yao et al., 2023; Zhao et al., 2023; Zhang et al., 2024; Xia et al., 2024) underscores the importance of thought decomposition and structured reasoning frameworks in improving both the accuracy and interpretability of LLM outputs. In particular, the intermediate steps of CoT can make the model's output easier to interpret and evaluate (Yu et al., 2023), which is valuable for tasks requiring high accountability, such as biomedical claim verification. Moreover, Wang et al. (2022) proposed the self-consistency method, which enhances the reliability of the results by sampling diverse CoT generations for each sample and selecting the most consistent conclusions among them. Weng et al. (2022) introduced backward verification to complement forward CoT reasoning, allowing self-verification of conclusions derived from different CoT paths to identify the most accurate CoT generations for specific tasks.

Pre-trained Language Models for Biomedical NLP In various NLP tasks, pre-trained language models (PLMs) are effectively applied to medical text processing. (Liang et al., 2023; Liang and Sonntag, 2024) investigated building German clinical entity extraction system based on German PLMs in low-data setting. More recent studies have explored the potential applications of PLMs in clinical practice, such as building clinical entity extraction system without in-domain training data (Liang and Sonntag, 2024), ranging from transfer learning in summarizing radiology reports (Liang et al., 2022) to real-time radiology reporting (Elkassem and Smith, 2023; Jeblick et al., 2024) with PLMs. Datta et al. (2024) leveraged PLMs for automatic eligibility criteria from free text clinical trial protocol to facilitate trial enrollment and evaluation. (Liu et al., 2024) demonstrated the potential of automated verification of scientific claims with LLMs using retrieval-augmented strategies that exploit open resources such as PubMed.

Sivarajkumar et al. (2024) highlighted the effectiveness of different prompting strategies, including zero-shot and few-shot, for clinical information extraction, while Tang et al. (2023) found that LLMs still struggle to summarize medical evidence in longer textual contexts by evaluating LLM-generated summaries focused on six clinical domains. Moreover, LLMs have been shown to enhance the diagnostic accuracy of general radiologists in cardiac imaging, highlighting their value as a diagnostic support tool (Cesur et al., 2024). Rao et al. (2023) also underscored the potential of LLMs to assist healthcare professionals in diagnostic decision-making. Studies from Benary et al. (2023) suggest that LLMs are not yet suitable for routine use in personalized clinical decision-making in oncology, they show promise as a complementary tool, such as selecting relevant biomedical literature to support evidence-based, personalized treatment decisions and offering unique strategies not identified by experts. However, further research is necessary to evaluate their integration into clinical workflows effectively (Verlingue et al., 2024).

7 Conclusion

In summary, our approach structures the complex NLI process into a sequential framework. The process begins with semantic grounding, where the model activates contextual understanding based on the statement to be verified. Next, the model identifies the relevant evidence from the premise data, where the model compares the information in the statement with the extracted evidence. After this evaluation, the model is asked to draw a conclusion and predict the logical relationship between the statement and the evidence. In the context of validating biomedical claims based on long and nuanced documents, the semantic grounding and evidence-based evaluation steps help LLMs perform sub-tasks with greater precision in contrast to the abstract nature of logical relationship prediction. We find that decomposition reduces ambiguity in textual understanding, making the LLM's responses less sensitive to specific wording, as long as the sub-tasks are clearly defined in prompting instructions. For example, the *claim comprehension* step only interprets key terms, while the *evidence analysis* focuses on comparing the statement and the evidence to identify relevant data points. This approach can also effectively minimize the need for extensive prompt engineering.

Future work In high-stakes areas such as medical decision-making, allowing LLMs to make decisions raises critical concerns about accountability and trustworthiness (Elkassem and Smith, 2023; Jeblick et al., 2024). Integrating a feedback-driven loop would support the development of collaborative systems that balance the responsibility for decision making between users and LLMs. This balance is particularly important in high-stakes domains where trust and accountability are essential.

Limitations

Our focus in this work has primarily been on the reasoning capabilities of models when relevant source documents are provided, with pre-retrieved documents used in the evaluation data. However, for open-ended cases, we would need to incorporate a retrieval pipeline to limit the candidate documents to a manageable scale, as otherwise, the process of evidential evaluation could become too time-consuming. Additionally, due to time constraints, we did not compare many different CoT methods. Some approaches, such as generating multiple responses and applying voting heuristics, could offer more reliable results but are computationally expensive. We opted for the most intuitive and effective method, focusing on the 4-step prompting framework. Furthermore, While LLMs demonstrate significant improvements in generating evaluations within 4-step strategy and after SFT, the degree of autonomy granted to these models should be further explored to align with specific user preferences and the application domain.

Acknowledgments

This work is funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), and by the Endowed Chair of Applied AI at the University of Oldenburg. We gratefully acknowledge the support provided by a grant from Accenture Labs.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

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 (EMNLP)*. Association for Computational Linguistics.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Turay Cesur, Yasin Celal Gunes, Eren Camur, and Mustafa Dagli. 2024. Empowering radiologists with chatgpt-4o: Comparative evaluation of large language models and radiologists in cardiac cases. *medRxiv*, pages 2024–06.

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.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Asser Abou Elkassem and Andrew D Smith. 2023. Potential use cases for chatgpt in radiology reporting. *American Journal of Roentgenology*, 221(3):373–376.

Aryo Gema, Giwon Hong, Pasquale Minervini, Luke Daines, and Beatrice Alex. 2024. Edinburgh clinical NLP at SemEval-2024 task 2: Fine-tune your model unless you have access to GPT-4. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1894–1904, Mexico City, Mexico. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2024. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 34(5):2817–2825.

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*.

Mael 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)*, pages 1947–1962, Mexico City, Mexico. 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

Siting Liang, Mareike Hartmann, and Daniel Sonntag. 2023. Cross-domain german medical named entity recognition using a pre-trained language model and unified medical semantic types. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 259–271.

Siting Liang, Klaus Kades, Matthias A Fink, Peter M Full, Tim F Weber, Jens Kleesiek, Michael Strube, and Klaus Maier-Hein. 2022. Fine-tuning bert models for summarizing german radiology findings. *ClinicalNLP 2022*, page 30.

Siting Liang and Daniel Sonntag. 2024. Building a german clinical named entity recognition system without in-domain training data. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 70–81.

Hao Liu, Ali Soroush, Jordan G Nestor, Elizabeth Park, Betina Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, et al. 2024. Retrieval augmented scientific claim verification. *JAMIA open*, 7(1):ooae021.

OpenAI. 2024. Openai models. available online. https://platform.openai.com/docs/models/overview. (accessed in November 2024).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dreyer, and Marc D Succi. 2023. Assessing the utility of chatgpt throughout the entire clinical workflow: development and usability study. *Journal of Medical Internet Research*, 25:e48659.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Unsloth. 2024. unsloth.ai. available online. https://docs.unsloth.ai/. (accessed in November 2024).

Loïc Verlingue, Clara Boyer, Louise Olgiati, Clément Brutti Mairesse, Daphné Morel, and Jean-Yves Blay. 2024. Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. *The Lancet Regional Health–Europe*, 46.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *Preprint*, arXiv:2404.15676.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv. org/pdf/2305.10601. pdf*.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *Preprint*, arXiv:2310.04959.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *arXiv preprint arXiv:2406.09136*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Appendix

Size of Applied Models Table 7 provides a comparison of model size and initial context window length. The model size of the open source LLMs is limited to 14 billion parameters. All models are the instruct fine-tuned version.

Model	Version	Context Window	Parameters
GPT3.5	gpt-3.5-turbo-0125	16K	175B
GPT-4o-mini	gpt-4o-mini-2024-07-18	128K	?
Phi3.5-3.6B	Phi-3.5-mini-instruct	128K	3.6B
Mistral-7b	mistral-7b-instruct-v0.3	32K	7B
Llama3.1-8B	Meta-Llama-3.1-8B-Instruct	128K	8B
Gemma2-9B	gemma-2-9b-bnb-it	8K	9B
Mistral-12B	Mistral-Nemo-Instruct-2407	1024K	12B
Phi3-14B	Phi-3-medium-4k-instruct	4K	14B

Table 7: List of low-cost GPT models and lightweight open-source LLMs used in our experiments.

Comparisons of generations of different models with different prompting strategies. In particular, Table 10-15 illustrates the enhancement of lightweights LLMs in analyzing the statement based on the provided data (The first step of 2-step CoT and the second step - *evidence analysis* in the 4-step framework) for the following example shown in Table 8.

Fine-tuned results are obtained after fine-tuning with the GPT-4o-mini augmented training samples. Fine-tuning provides the LLMs with reliable analysis patterns for the NLI4CT tasks, thereby increasing the reliability of small-scale models. We utilize the *FastLanguageModel* modules of **unsloth**[1] library to accelerate the SFT fine-tuning in our experiments. **SFT** in the tables represents the response in the second step (EVIDENCE ANALYSIS) by the model fine-tuned with GPT-4o-mini generated responses.

[1] https://github.com/unslothai/unsloth

Statement: ’Only 2 patients in the primary trial did not have Recurrence-free Survival’
Clinical Trial Data: Primary trial: - Outcome Measurement: - Recurrence-free Survival - Recurrence-free survival curves will be plotted for subjects treated with stage I and II disease. - Time frame: Time from the start of treatment to recurrence, second malignancy, or death as a first event, assessed up to 3 years - Results 1: - Arm/Group Title: Treatment (Chemotherapy With or Without Maintenance Therapy) - Arm/Group Description: SYSTEMIC CHEMOTHERAPY: Patients receive cyclophosphamide IV over 1 hour and paclitaxel IV over 3 hours on day 1. Treatment repeats every 14 days for 6 courses in the absence of disease progression or unacceptable toxicity. - MAINTENANCE THERAPY (Her-2 neu positive patients): Patients receive trastuzumab IV over 30 minutes on day 1. Treatment repeats every 14 days for 5 courses and then every 21 days for 14 courses in the absence of disease progression or unacceptable toxicity. - cyclophosphamide, paclitaxel, trastuzumab: Given IV - Overall Number of Participants Analyzed: 100 - Measure Type: Number - Unit of Measure: percentage of subjects 98 (92.2 to 99.5)
Label: Entailment

Table 8: On exmpale instance of ’Entailment’ class from NLI4CT test set.

GPT-4o-mini in *evidence analysis* step: "To evaluate the claim "Only 2 patients in the primary trial did not have Recurrence-free Survival," we first need to extract and clarify the relevant data points from the provided clinical trial data and then analyze the claim against those data points. 1. Relevant Data Points Extracted and Clarified: - **Total Participants Analyzed**: 100 - **Recurrence-free Survival Result**: 98 participants (which equates to the percentage of subjects achieving RFS) achieved recurrence-free survival. - **Percentage of Subjects with RFS**: 98% (confidence interval of 92.2% to 99.5%) 2. Evaluation of the Claim against Data Points: 1. **Total number of patients**: The total number of participants in the trial is 100. - **Evaluation**: The claim mentions "only 2 patients" not achieving RFS. This would imply that a total of 98 out of 100 patients experienced RFS as stated in the trial data. 2. **Number of patients with Recurrence-free Survival**: According to the trial data, 98 patients achieved RFS. - **Evaluation**: This supports the claim, as if 98 patients achieved RFS, then logically it follows that 2 patients did not achieve RFS (100 total participants - 98 with RFS = 2 without RFS). 3. **Effectiveness of Treatment**: The high percentage of 98% suggests a strong effectiveness of the treatment administered. - **Evaluation**: This supports the assertion in the claim that such a small number of patients (2 in this case) who did not achieve RFS indicates high treatment effectiveness. 4. **Definition of Recurrence-Free Survival**: RFS is defined as a period during which a patient remains free of any signs of disease recurrence. - **Evaluation**: The claim directly referenced RFS, which is consistent with the clinical outcome measures defined in the trial data. Conclusion: The claim that "Only 2 patients in the primary trial did not have Recurrence-free Survival" is consistent with the clinical trial data provided. The data indicates that out of 100 participants, 98 achieved RFS, meaning it is accurate to state that only 2 patients did not have RFS. Hence, the claim is validated by the data."
Predict: Entailment

Table 9: Evaluation on the above example made by GPT-4o-mini in the EVIDENCE ANALYSIS that leads to the accurate logical relation prediction.

Model	Approach	Prediction
Phi3.5-3.6B	**zero-shot 2-step:** "The statement claims that "Only 2 patients in the primary trial did not have Recurrence-free Survival," but the clinical trial data does not provide a specific number of patients without Recurrence-free Surv"	Contradiction
	zero-shot 4-step:"Recurrence-free Survival" (RFS) is the primary outcome measure, which in this context is the time until cancer recurrence, development of a second cancer, or death, whiche0ever comes first. - "Time frame" specifies that the RFS is assessed from the start of treatment up to 3 years, which is a standard period for evaluating long-term outcomes in oncology trials. - "Arm/Group Title: Treatment (Chemotherapy With or Without Maintenance Therapy)" refers to the two distinct treatment strategies being compared in the trial: one with systemic chemotherapy alone and the other with chemotherapy followed by maintenance therapy for Her-2 neu positive patients. - "Arm/Group Description" provides details on the chemotherapy regimen for the systemic treatment arm, including the drugs used (cyclophosphamide and paclitaxel), the administration method (IV), and the treatment schedule (every 14 days for 6 courses). - "MAINTENANCE THERAPY (Her-2 neu positive patients)" specifies the additional treatment for patients with Her-2 neu positive cancer, which includes the drug trastuzumab, also administered IV, with a different treatment schedule (every 14 days for 5 courses, then every 21 days for 14 courses). - "cyclophosphamide, paclitaxel, trastuzumab: Given IV" confirms the intravenous administration of the drugs used in the trial. - "Overall Number of Participants Analyzed: 100"	Entailment
	SFT: " To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" based on the provided clinical trial data, we can analyze the relevant information as follows: 1. **Outcome Measurement**: The primary outcome measure is Recurrence-free Survival (RFS), which assesses the time from the start of treatment to recurrence, second malignancy, or death. 2. **Time Frame**: The assessment period for RFS is up to 3 years, indicating a focus on short- to medium-term outcomes. 3. **Treatment Arm**: The trial includes two treatment groups: - **Chemotherapy With or Without Maintenance Therapy**: This group receives systemic chemotherapy (cyclophosphamide and paclitaxel) and, for Her-2 neu positive patients, maintenance therapy (trastuzumab). - **Maintenance Therapy Only**: This group receives maintenance therapy (trastuzumab) only. 4. **Overall Number of Participants Analyzed**: A total of 100 participants were included in the trial. 5. **Results**: The reported outcome indicates that 98% of participants achieved Recurrence-free Survival. This translates to: - 98% of 100 participants = 98 participants had Recurrence-free Survival. - Therefore, 2 patients (100 - 98) did not have Recurrence-free Survival. Conclusion: The statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" is supported by the data. The evidence shows that 98 participants achieved Recurrence-free Survival, confirming that indeed only 2 patients did not experience this outcome. Thus, the statement can be logically inferred from the provided clinical trial data."	Entailment

Table 10: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Phi3.5-3.6B.

Model	Approach	Prediction
Mistral-7B	**zero-shot 2-step:** The statement does not logically follow from the provided clinical trial data. The data only provides the percentage of subjects (98 out of 100) who did not experience a recurrence, but it does not specify the number of subjects who did experience a recurrence. Therefore, it is not possible to determine from this data whether only 2 patients in the primary trial did not have Recurrence-free Survival."	Contradiction
	zero-shot 4-step:" Based on the provided clinical trial data, the claim "Only 2 patients in the primary trial did not have Recurrence-free Survival" can be evaluated as follows: 1. Outcome Measurement: Recurrence-free Survival The claim and the data both refer to the same outcome measure. 2. Time frame: Time from the start of treatment to recurrence, second malignancy, or death as a first event, assessed up to 3 years The time frame in the claim is consistent with the time frame in the data (up to 3 years). 3. Results 1: - Arm/Group Title: Treatment (Chemotherapy With or Without Maintenance Therapy) - Arm/Group Description: SYSTEMIC CHEMOTHERAPY: Patients receive cyclophosphamide IV and paclitaxel IV. MAINTENANCE THERAPY (Her-2 neu positive patients): Patients receive trastuzumab IV. - Overall Number of Participants Analyzed: 100 - Measure Type: percentage of subjects - Unit of Measure: percentage The data shows that 98 out of 100 participants (92.2 to 99.5%) did not have recurrence-free survival. This suggests that the majority of participants in the trial did not experience a recurrence of their disease during the follow-up period. However, the claim does not specify which arm or group within the trial the 2 patients without recurrence-free survival belong to. In conclusion, while the data supports the claim that the majority of participants did not experience a recurrence of their disease, the claim is not fully accurate as it does not specify which arm or group the 2 patients without recurrence-free survival belong to. A more detailed analysis would require information"	Contradiction
	SFT: "To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival," we can analyze the provided clinical trial data as follows: 1. **Outcome Measurement**: The primary outcome of the trial is Recurrence-free Survival (RFS), which is defined as the time from the start of treatment to recurrence, second malignancy, or death, assessed up to 3 years. 2. **Results Overview**: The trial included a total of 100 participants, with 98 achieving recurrence-free survival, resulting in a percentage of 98% (with a confidence interval of 92.2% to 99.5%). This indicates that a significant majority of participants remained free from disease recurrence. 3. **Interpretation of the Statement**: The statement claims that only 2 patients did not have recurrence-free survival. Given that 98 out of 100 participants achieved this outcome, it can be inferred that the remaining 2 patients did not achieve recurrence-free survival. 4. **Conclusion**: The evidence from the trial data supports the statement, confirming that only 2 patients did not have recurrence-free survival, as they are the 2 out of 100 participants who did not achieve this outcome. Thus, the statement can be accurately inferred from the provided data."	Entailment

Table 11: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Mistral-7B.

Model	Approach	Prediction
Llama3.1-8B	**zero-shot 2-step:** " Based on the provided clinical trial data, the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" does not logically follow from the given information. The statement in question implies that the outcome of interest is the number of patients without Recurrence-free Survival, which is not explicitly mentioned in the provided data. The data only mentions the outcome of Recurrence-free Survival, not the absence of it."	Contradiction
	zero-shot 4-step: " Based on the clinical trial data, the following relevant data points are identified: * **Outcome Measure**: Recurrence-free Survival * **Time Frame**: Up to 3 years * **Number of Participants Analyzed**: 100 * **Measure Type**: Number * **Unit of Measure**: Percentage of subjects Evaluating the claim: * The claim states that "Only 2 patients in the primary trial did not have Recurrence-free Survival". This statement seems to be inconsistent with the data, as the trial reported a 98% (92.2-99.5) recurrence-free survival rate, indicating that a significant portion of the patients did experience recurrence-free survival. * The claim suggests that only 2 patients out of the total 100 participants did not achieve recurrence-free survival, which contradicts the overall recurrence-free survival rate of 98% reported in the trial. * The trial's reported results do not support the claim that only 2 patients did not experience recurrence-free survival."	Contradiction
	SFT: " To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival," we will analyze it in relation to the provided clinical trial data. Evaluation of the Statement: 1. **Outcome Measurement**: The primary trial measures Recurrence-free Survival, which indicates the duration patients remain free from disease recurrence following treatment. 2. **Patient Cohort**: The statement refers to a specific group of patients within the trial, focusing on those who did not achieve recurrence-free survival. 3. **Results Data**: - **Total Participants Analyzed**: 100 patients. - **Outcome Measure**: The percentage of participants achieving recurrence-free survival is reported as 98% (with a confidence interval of 92.2% to 99.5%). 4. **Interpreting the Results**: - If 98% of patients achieved recurrence-free survival, this means that: - 98% of $100 = 98$ patients did have recurrence-free survival. - Consequently, the number of patients who did not achieve recurrence-free survival would be: $100 - 98 = 2 \text{ patients}$ Conclusion: The statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" can be logically inferred from the provided data, confirming that indeed, 2 out of the 100 participants did not achieve recurrence-free survival, supporting the accuracy of the claim."	Entailment

Table 12: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Llama3.1-8B.

Model	Approach	Prediction
Gemma2-9B	**zero-shot 2-step:** "The statement **does not logically follow** from the provided clinical trial data. Here's why: * **Missing Data:** The data only provides the overall percentage of subjects (98%) who experienced Recurrence-free Survival. It doesn't specify the exact number of patients who did or did not experience recurrence. * **Potential for Interpretation:** While the high percentage suggests that most patients did experience Recurrence-free Survival, it doesn't definitively confirm that only two patients did not. To determine if the statement is true, we would need the raw data on individual patient outcomes (e.g., a table showing whether each patient experienced recurrence or not). "	Contradiction
	zero-shot 4-step:"Let's break down the claim and compare it to the provided data: **Claim:** 'Only 2 patients in the primary trial did not have Recurrence-free Survival' **Data Points:** * **Primary Trial:** Yes, the data refers to a primary trial. * **Outcome Measurement:** Recurrence-free Survival is the primary outcome. * **Time Frame:** The trial assesses Recurrence-free Survival up to 3 years. * **Overall Number of Participants Analyzed:** 100 * **Results 1:** The data shows 98 out of 100 (92.2 to 99.5%) of patients in the treatment arm experienced Recurrence-free Survival. **Evaluation:** The claim is **partially supported** by the data. * **Support:** The data clearly states that 98 out of 100 patients in the treatment arm had Recurrence-free Survival. This means only 2 patients did *not* experience Recurrence-free Survival. * **Missing Information:** The claim doesn't specify which treatment arm it refers to. We need to know if this refers to the chemotherapy with or without maintenance therapy. **Important Considerations:** * **Confidence Intervals:** The data provides a confidence interval (92.2 to 99.5%). This means we are 95% confident that the true percentage of patients with Recurrence-free Survival in the treatment arm falls within this range. * **Context:** To fully evaluate the claim, we need more information about the control group (if there was one) and the specific types of cancer being treated. "	Entailment
	SFT:" To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" based on the provided clinical trial data, we can analyze the relevant information as follows: 1. **Outcome Measurement**: The primary trial focused on measuring Recurrence-free Survival (RFS), which is a critical endpoint in cancer trials. 2. **Time Frame**: The assessment of RFS was conducted over a period of up to 3 years from the start of treatment, providing a comprehensive view of the treatment's effectiveness. 3. **Results Analysis**: - The trial included a total of **100 participants** analyzed. - The statement indicates that **2 patients** did not achieve RFS. - This translates to **98%** of participants having RFS, as shown in the results (98% with a confidence interval of 92.2% to 99.5%). 4. **Conclusion**: The statement is supported by the data, confirming that only 2 out of 100 participants did not achieve recurrence-free survival. This indicates a high level of effectiveness of the treatment in preventing disease recurrence among the majority of patients. In summary, the evidence from the trial data clearly supports the statement that only 2 patients did not have recurrence-free survival, highlighting the treatment's success in maintaining disease-free status in the majority of participants."	Entailment

Table 13: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Gemma2-9B.

Model	Approach	Prediction
Mistral-12B	**zero-shot 2-step:** "Based on the provided clinical trial data, the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" is **not logically supported** by the data. Here's why: 1. The clinical trial data reports that 98 out of 100 participants had Recurrence-free Survival, which is 92.2% to 99.5% (with 95% confidence interval). 2. This means that, at most, 2 out of 100 participants did not have Recurrence-free Survival, which is consistent with the reported percentage. 3. However, the statement uses "only 2 patients" which implies that the number is exact and not an estimate. The data provided does not confirm this exact number; it only gives a range (92.2% to 99.5%) within which the number of patients with Recurrence-free Survival falls. "	Contradiction
	zero-shot 4-step:"Based on the provided clinical trial data, let's evaluate the claim "Only 2 patients in the primary trial did not have Recurrence-free Survival": 1. **Number of participants in the trial**: The claim mentions "primary trial," and the data provides the overall number of participants analyzed as 100. This matches the context of the claim, as it refers to the primary trial. - Data point: 100 participants - Claim: Applicable (primary trial) 2. **Outcome measure**: The claim focuses on Recurrence-free Survival (RFS), which is also the primary outcome measurement mentioned in the trial data. - Data point: Recurrence-free Survival - Claim: Applicable (RFS is the outcome of interest) 3. **Time frame**: The claim does not specify a time frame, but the trial data indicates that RFS is assessed up to 3 years. Without a specified time frame in the claim, we cannot directly evaluate this aspect. - Data point: Up to 3 years - Claim: Not specified, so cannot be evaluated 4. **Number of patients without RFS**: The claim states "Only 2 patients... did not have Recurrence-free Survival." The trial data does not provide the exact number of patients without RFS, but it does mention the overall percentage of subjects with RFS. - Data point: 98% (92.2 to 99.5) of subjects had RFS - Claim: The data suggests that 2% (0.5 to 7.8) of subjects did not have RFS, which is consistent with the claim's statement of "Only 2 patients." 5. **Arm/Group**: The claim does not specify any particular treatment arm or group, so we cannot directly evaluate this aspect."	Entailment
	SFT: "To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival," we can analyze the provided clinical trial data as follows: 1. **Outcome Measurement**: The trial measures Recurrence-free Survival, which is defined as the time from the start of treatment to recurrence, second malignancy, or death. 2. **Total Participants**: The trial included 100 participants. 3. **Results**: The statement claims that only 2 patients did not have Recurrence-free Survival. This implies that 98 patients did achieve Recurrence-free Survival. 4. **Confidence Interval**: The reported percentage of subjects with Recurrence-free Survival is 98%, with a confidence interval of (92.2 to 99.5). This indicates a high level of confidence in the outcome. 5. **Conclusion**: Based on the data, it can be inferred that the statement is accurate. If 98 out of 100 participants achieved Recurrence-free Survival, then the number of patients who did not achieve this outcome is indeed 2, confirming the statement.""	Entailment

Table 14: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Mistral-12B .

Model	Approach	Prediction
Phi3-14B	**zero-shot 2-step:** " The statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" does not logically follow from the clinical trial data provided. The data indicates that 98% (92.2 to 99.5) of the subjects had recurrence-free survival, which means that 2% (100 - 98%) of the subjects did not have recurrence-free survival. However, the data does not specify the exact number of patients who did not have recurrence-free survival."	Contradiction
	zero-shot 4-step:" Relevant data points from the clinical trial data: 1. Outcome Measurement: Recurrence-free Survival 2. Time frame: Time from the start of treatment to recurrence, second malignancy, or death as a first event, assessed up to 3 years 3. Arm/Group Title: Treatment (Chemotherapy With or Without Maintenance Therapy) 4. Arm/Group Description: SYSTEMIC CHEMOTHERAPY and MAINTENANCE THERAPY (Her-2 neu positive patients) 5. Overall Number of Participants Analyzed: 100 6. Measure Type: Number 7. Unit of Measure: percentage of subjects 8. Results 1: 98 (92.2 to 99.5) Evaluation of the claim: The claim states that "Only 2 patients in the primary trial did not have Recurrence-free Survival." Based on the clinical trial data provided, the results show that 98% of the subjects (92.2 to 99.5) had Recurrence-free Survival. This means that 2% of the subjects did not have Recurrence-free Survival. Given that there were 100 participants in the trial, 2% of the participants would be 2 participants. Therefore, the claim is consistent with the clinical trial data, as it accurrightly states that only 2 patients in the primary trial did not have Recurrence-free Survival."	Entailment
	SFT: " To evaluate the statement "Only 2 patients in the primary trial did not have Recurrence-free Survival," we can analyze the provided clinical trial data as follows: 1. **Outcome Measurement**: The primary outcome of the trial is Recurrence-free Survival (RFS), which is defined as the time from the start of treatment to the occurrence of recurrence, a second malignancy, or death. 2. **Treatment Description**: The trial involves systemic chemotherapy, with specific regimens for Her-2 neu positive patients, including cyclophosphamide, paclitaxel, and trastuzumab. 3. **Time Frame**: The assessment of RFS is conducted over a period of up to 3 years from the start of treatment. 4. **Overall Number of Participants Analyzed**: The trial includes a total of 100 participants. 5. **Results**: The data indicates that 98% of the participants achieved recurrence-free survival, with a confidence interval of 92.2% to 99.5%. Evaluation of the Statement: - **Claim**: "Only 2 patients in the primary trial did not have Recurrence-free Survival." - **Analysis**: - If 98% of the 100 participants achieved RFS, this means that 2% did not. - Calculating the number of patients who did not achieve RFS: - 2% of 100 participants = 2 patients. Conclusion: The statement "Only 2 patients in the primary trial did not have Recurrence-free Survival" is supported by the data. The evidence indicates that 2 patients out of 100 did not achieve RFS, confirming the accuracy of the claim."	Entailment

Table 15: Comparison between the model responses made in the first step of zero-shot 2-step framework and the second step (*evidence analysis*) in the zero-shot 4-step framework, made by model Phi3-14B.