Explainable Biomedical Claim Verification with Large Language Models

Siting Liang[1], Daniel Sonntag[1,2]

[1]German Research Center for Artificial Intelligence, Germany
[2]University of Oldenburg, Germany

Abstract

Verification of biomedical claims is critical for healthcare decision-making, public health policy and scientific research. We present an interactive biomedical claim verification system by integrating LLMs, transparent model explanations, and user-guided justification. In the system, users first retrieve relevant scientific studies from a persistent medical literature corpus and explore how different LLMs perform natural language inference (NLI) within task-adaptive reasoning framework to classify each study as "Support," "Contradict," or "Not Enough Information" regarding the claim. Users can examine the model's reasoning process with additional insights provided by SHAP values that highlight word-level contributions to the final result. This combination enables a more transparent and interpretable evaluation of the model's decision-making process. A summary stage allows users to consolidate the results by selecting a result with narrative justification generated by LLMs. As a result, a consensus-based final decision is summarized for each retrieved study, aiming safe and accountable AI-assisted decision-making in biomedical contexts. We aim to integrate this explainable verification system as a component within a broader evidence synthesis framework to support human-AI collaboration.

Keywords

Biomedical Claim Verification, Large Language Models, Natural Language Inference, Explainable AI

1. Introduction

Automated biomedical claim verification systems aim to assist clinicians and researchers in combating the potential harm of misinformation in the healthcare domain. These systems verify claims related to treatments, clinical trial outcomes, and other medical assertions by synthesizing evidence from clinical trial data and scientific literature. Biomedical claim verification involves assessing the veracity of such claims through relevant studies, ensuring reliable conclusions for critical decision-making [1, 2, 3, 4]. Research in biomedical claim verification has focused on utilizing advanced natural language processing (NLP) techniques. Fine-tuned natural language inference (NLI) models have been widely adopted for this task [5]. These models typically follow a standard pipeline: (1) retrieve relevant studies using the claim as query, (2) process the claim and retrieved evidence using a language model in either a fine-tuned or in-context learning setup designed for NLI task [6, 7]. The NLI task requires determining the logical relationship between two pieces of text: a premise (in our case, scientific studies) and the biomedical claim to be verified. The task involves classifying this relationship into one of three categories: Support, Contradict and No Enough Information. Related examples from Wadden et al. [1] are shown in Table 1. In the scientific and medical domains, NLI models are required to process long and complex documents while also a deep understanding of biomedical knowledge to interpret specific terminologies [8, 6, 7]. In particular, scientific studies often contain complex statistical information and precise measurements that must be interpreted accurately to avoid errors in claim verification [9]. Large language models (LLMs) offer promising potential to address these challenges [10]. Their effectiveness depends on two critical factors: the size of the model and the suitability of the prompts designed for specific tasks [11, 12, 13].

Real-world applications of LLMs, especially in healthcare and scientific domains, demand high levels of transparency, interpretability, and trustworthiness due to the high-stakes nature of decisions

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

siting.liang@dfki.de (S. Liang); daniel.sonntag@dfki.de (D. Sonntag)
0009-0000-0146-581X (S. Liang); 0000-0002-8857-8709 (D. Sonntag)

Claim	Most Relevant Study	Relation
76-85% of people with severe mental disorder receive no treatment in low and middle income countries.	... RESULTS The prevalence of having any WMH-CIDI/DSM-IV disorder in the prior year varied widely, from 4.3% in Shanghai to 26.4% in the United States.... Although disorder severity was correlated with probability of treatment in almost all countries, 35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview. Due to the high prevalence of mild and subthreshold cases, the number of those who received treatment far exceeds the number of untreated serious cases in every country.	Support / Entailment
10-20% of people with severe mental disorder receive no treatment in low and middle income countries.		Contradict

Table 1
Examples of biomedical claim verification illustrating the logical relationship between two claims and the retrieved most relevant study. Phrases highlighted are the critical statistic information that determines the logical relation between the claims and study.

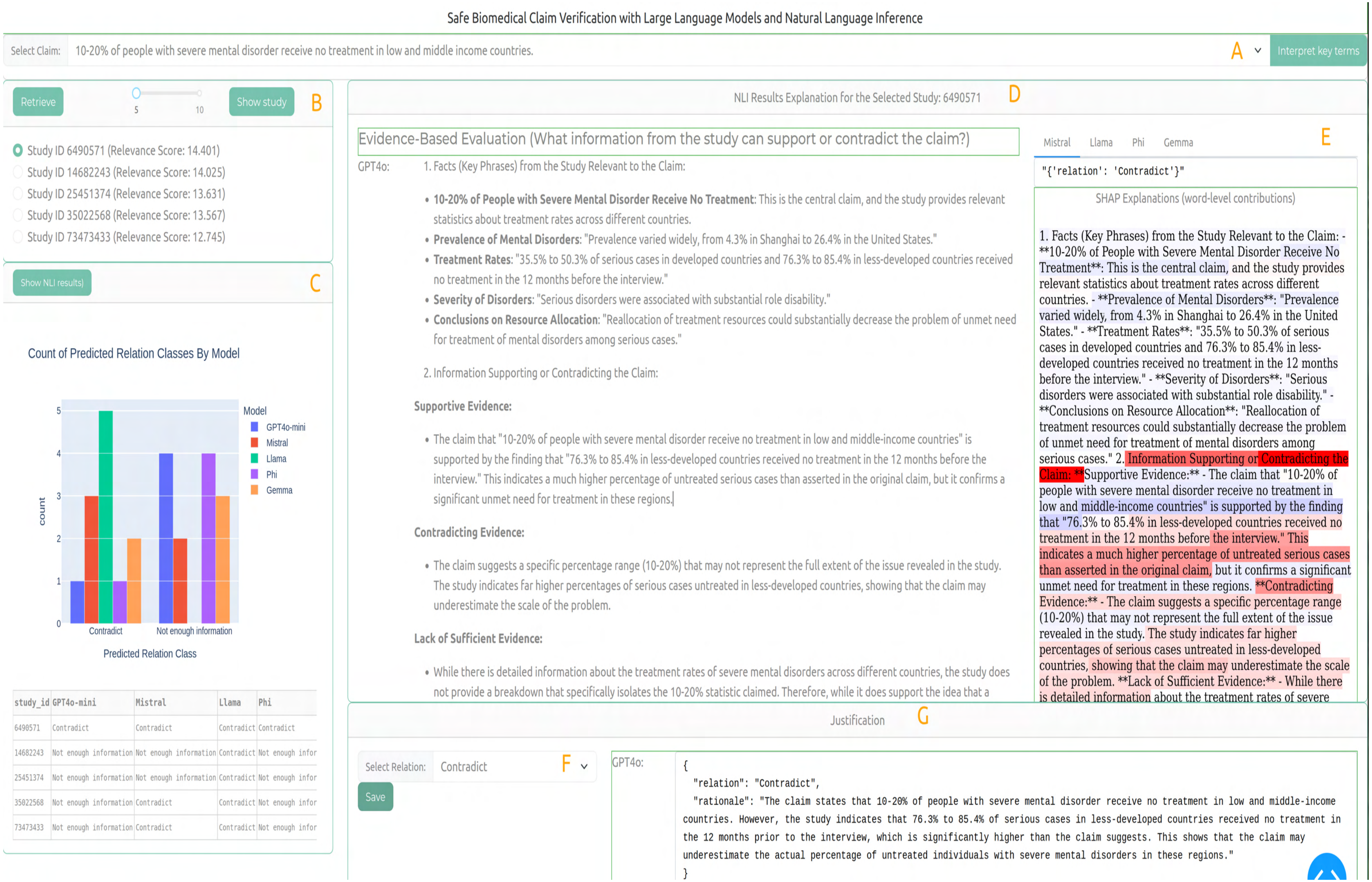
[14, 15, 16]. In this work, we propose an interactive biomedical claim verification system as part of the *No-IDLE* [17] and *No-IDLE meet ChatGPT* [18] projects about interactive deep learning and LLMs. The primary functionality of the system is to assist users in validating claims by leveraging the strengths of LLM-based verification while ensuring a transparent and reliable decision-making process. The system builds on the Chain of Evidential Natural Language Inference (**CoENLI**) framework, which enables LLMs to generate evidence-based rationales before arriving at a final relation classification. To evaluate the framework, we use two relevant biomedical benchmarks, demonstrating that (**CoENLI**) significantly improves LLM accuracy and outperforms the general Chain of Thought (CoT) approach [19]. The evaluation results demonstrate that by explicitly outlining the evidence-based reasoning process, **CoENLI** enhances both the interpretability and reliability of claim verification.

To further enhance the interpretability of the system, we integrate Shapley Additive exPlanations (SHAP) saliency maps [20], which highlights the word-level contributions within the generated rationales. This technique provides a deeper understanding of how LLMs weigh specific evidence for arriving at final conclusion. In addition, the system employs different LLMs to provide users with a comparative analysis of results and reasoning. By enabling users to review different perspectives and reasoning outputs, the system fosters a nuanced understanding of the claim verification process, ultimately increasing trustworthiness of LLMs in real-world applications. Finally, users select the most appropriate classification as the final decision after reviewing model-generated rationales and explanations. Our main contribution in this work is the development of an iterative human-AI collaboration workflow that ensures transparency, accountability, and adaptability to individual expert knowledge. By leveraging LLMs with **CoENLI**, the system delivers transparent evidence-based evaluations while incorporating SHAP saliency explanations. Additionally, comparative insights from multiple LLMs further enhance understandability and reliability. Together, these innovations advance automatic biomedical claim verification towards greater confidence and usability in the process.

2. Explainable Biomedical Claim Verification System

2.1. Overview

The system comprises several interactive components to combine the strengths of advanced language models-driven analysis and user control. Figure 1 provides an overview of our system's user interface. Users initiate the verification process by selecting a claim to investigate (A) and retrieving relevant studies from database of scientific literatures using BM25 algorithm [21] to the claim of being assessed (B). The system employs multiple LLMs within the **CoENLI** framework to evaluate the relationship between the claim and each retrieved study (C).


Safe Biomedical Claim Verification with Large Language Models and Natural Language Inference
A
Interpret key terms
Retrieve
5
10
Show study
B
Show NLI results)
C
Count of Predicted Relation Classes By Model
count
0
1
2
3
4
5
Contradict
Not enough information
Predicted Relation Class
Model
GPT4o-mini
Mistral
Llama
Phi
Gemma
NLI Results Explanation for the Selected Study: 6490571
D
Mistral
Llama
Phi
Gemma
E
SHAP Explanations (word-level contributions)
Justification
G
F
Save

Select Claim: 10-20% of people with severe mental disorder receive no treatment in low and middle income countries.

● Study ID 6490571 (Relevance Score: 14.401)

○ Study ID 14682243 (Relevance Score: 14.025)

○ Study ID 25451374 (Relevance Score: 13.631)

○ Study ID 35022568 (Relevance Score: 13.567)

○ Study ID 73473433 (Relevance Score: 12.745)

study_id	GPT4o-mini	Mistral	Llama	Phi
6490571	Contradict	Contradict	Contradict	Contradict
14682243	Not enough information	Not enough information	Contradict	Not enough infor
25451374	Not enough information	Not enough information	Contradict	Not enough infor
35022568	Not enough information	Contradict	Contradict	Not enough infor
73473433	Not enough information	Contradict	Contradict	Not enough infor

Evidence-Based Evaluation (What information from the study can support or contradict the claim?)

GPT4o:

1. Facts (Key Phrases) from the Study Relevant to the Claim:

- **10-20% of People with Severe Mental Disorder Receive No Treatment**: This is the central claim, and the study provides relevant statistics about treatment rates across different countries.
- **Prevalence of Mental Disorders**: "Prevalence varied widely, from 4.3% in Shanghai to 26.4% in the United States."
- **Treatment Rates**: "35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview."
- **Severity of Disorders**: "Serious disorders were associated with substantial role disability."
- **Conclusions on Resource Allocation**: "Reallocation of treatment resources could substantially decrease the problem of unmet need for treatment of mental disorders among serious cases."

2. Information Supporting or Contradicting the Claim:

Supportive Evidence:

- The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions.

Contradicting Evidence:

- The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the problem.

Lack of Sufficient Evidence:

- While there is detailed information about the treatment rates of severe mental disorders across different countries, the study does not provide a breakdown that specifically isolates the 10-20% statistic claimed. Therefore, while it does support the idea that a

"{'relation': 'Contradict'}"

1. Facts (Key Phrases) from the Study Relevant to the Claim: - **10-20% of People with Severe Mental Disorder Receive No Treatment**: This is the central claim, and the study provides relevant statistics about treatment rates across different countries. - **Prevalence of Mental Disorders**: "Prevalence varied widely, from 4.3% in Shanghai to 26.4% in the United States." - **Treatment Rates**: "35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." - **Severity of Disorders**: "Serious disorders were associated with substantial role disability." - **Conclusions on Resource Allocation**: "Reallocation of treatment resources could substantially decrease the problem of unmet need for treatment of mental disorders among serious cases." 2. Information Supporting or Contradicting the Claim: **Supportive Evidence:** - The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions. **Contradicting Evidence:** - The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the problem. **Lack of Sufficient Evidence:** - While there is detailed information about the treatment rates of severe

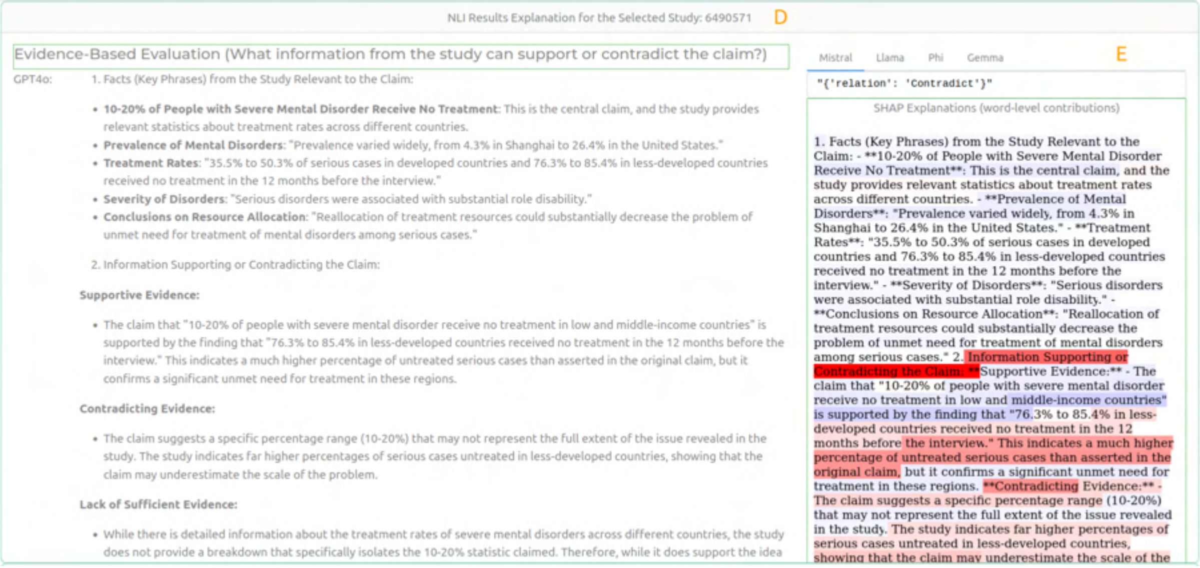
Select Relation: Contradict

GPT4o:

```
{
  "relation": "Contradict",
  "rationale": "The claim states that 10-20% of people with severe mental disorder receive no treatment in low and middle-income countries. However, the study indicates that 76.3% to 85.4% of serious cases in less-developed countries received no treatment in the 12 months prior to the interview, which is significantly higher than the claim suggests. This shows that the claim may underestimate the actual percentage of untreated individuals with severe mental disorders in these regions."
}
```

Figure 1: The biomedical claim verification system comprises several interactive components for the user study.

To enhance the transparency of the verification results, the interface displays model's analysis of evidence from the selected study that supports or contradicts the claim (D). To enhance understanding, we use SHAP values to highlight which parts of the rationales contributed most to the model's final decision, revealing the model's focus in drawing the conclusion (E) . This dual-layer interpretability, showing both how the model analyzes evidence and how it arrives at its final classification, provides users with a deeper understanding of the verification results (see Figure 2).



NLI Results Explanation for the Selected Study: 6490571

Evidence-Based Evaluation (What information from the study can support or contradict the claim?)

GPT4o:

1. Facts (Key Phrases) from the Study Relevant to the Claim:

- **10-20% of People with Severe Mental Disorder Receive No Treatment**: This is the central claim, and the study provides relevant statistics about treatment rates across different countries.
- **Prevalence of Mental Disorders**: "Prevalence varied widely, from 4.3% in Shanghai to 26.4% in the United States."
- **Treatment Rates**: "35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview."
- **Severity of Disorders**: "Serious disorders were associated with substantial role disability."
- **Conclusions on Resource Allocation**: "Reallocation of treatment resources could substantially decrease the problem of unmet need for treatment of mental disorders among serious cases."

2. Information Supporting or Contradicting the Claim:

Supportive Evidence:

- The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions.

Contradicting Evidence:

- The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the problem.

Lack of Sufficient Evidence:

- While there is detailed information about the treatment rates of severe mental disorders across different countries, the study does not provide a breakdown that specifically isolates the 10-20% statistic claimed. Therefore, while it does support the idea

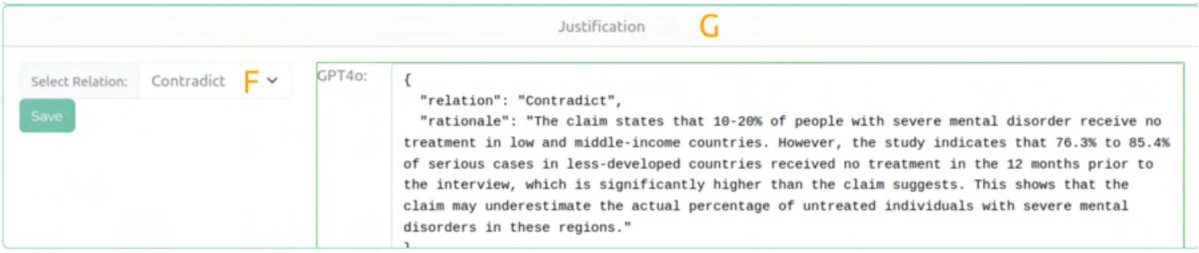
```
"{'relation': 'Contradict'}"
```

SHAP Explanations (word-level contributions)

1. Facts (Key Phrases) from the Study Relevant to the Claim: - **10-20% of People with Severe Mental Disorder Receive No Treatment**: This is the central claim, and the study provides relevant statistics about treatment rates across different countries. - **Prevalence of Mental Disorders**: "Prevalence varied widely, from 4.3% in Shanghai to 26.4% in the United States." - **Treatment Rates**: "35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." - **Severity of Disorders**: "Serious disorders were associated with substantial role disability." - **Conclusions on Resource Allocation**: "Reallocation of treatment resources could substantially decrease the problem of unmet need for treatment of mental disorders among serious cases." 2. Information Supporting or Contradicting the Claim: **Supportive Evidence:** - The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions. **Contradicting Evidence:** - The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the

Figure 2: Components D and E provide users dual-layer interpretability for a deeper understanding of verification results by combining evidence analysis with SHAP-based rationale highlighting.

After reviewing the generated rationales and saliency maps, if the users disagree with the initial classification result, they can adjust it (F), prompting the model to generate a concise justification for the updated classification (G) (see Figure 3).



Justification

Select Relation: Contradict

GPT4o:

```
{
  "relation": "Contradict",
  "rationale": "The claim states that 10-20% of people with severe mental disorder receive no treatment in low and middle-income countries. However, the study indicates that 76.3% to 85.4% of serious cases in less-developed countries received no treatment in the 12 months prior to the interview, which is significantly higher than the claim suggests. This shows that the claim may underestimate the actual percentage of untreated individuals with severe mental disorders in these regions."
}
```

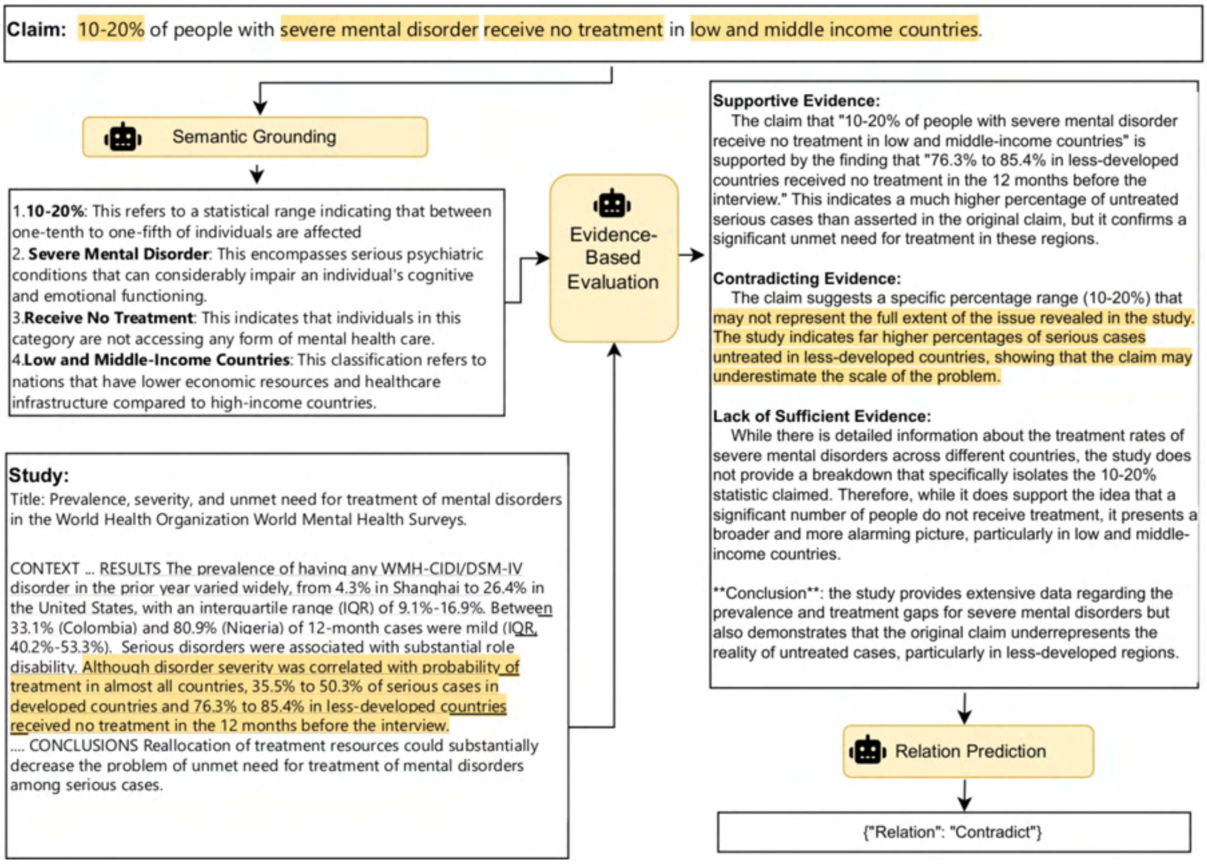
Figure 3: Components F and G comprise the summary stage, where users actively engage by adjusting the classification results and prompting the model to generate a final justification for the final decision.

This iterative approach allows for user involvement and enhanced trustworthiness in claim verification. More details about the **CoENLI** framework and SHAP values are explained in the subsections 2.2 and 2.3. The evaluation of the **CoENLI** framework and the choice of models are to be found in the section 3. Our study of the explainable claim verification system aims to assist human experts in making informed decisions by clearly presenting the evidence analyzed by LLMs while also enhancing transparency and providing comprehensive model explanations. The ultimate goal is to enable users to effectively assess the rationale behind the system's conclusions, fostering trust and facilitating collaboration in complex decision-making tasks.

2.2. Chain of Evidence-Based Natural Language Inference (CoENLI)

When prompting LLMs to complete reasoning tasks, breaking down complex reasoning tasks into simpler steps can be useful. Chain-of-Thought (CoT) strategies [19, 22], which provide exemplars of reasoning processes have demonstrated impressive performance across different reasoning benchmarks. Decomposition steps are useful for increasing the reliability of model generations [23]. Zhou et al. [24] noted that step-wise prompting require task-specific design for optimal performance. Inspired by prior works [24, 25], we propose **CoENLI** to refine the CoT reasoning in claim verification tasks including the following steps and Figure 4 illustrates the reasoning process within **CoENLI** with the example from Table 1.

- Semantic Grounding: A task instruction prompt contains the phrase *"Interpret the key terms in the claim"*. It activates specific semantic understanding of biomedical knowledge and logical patterns in LLMs, providing a contextual foundation for the subsequent reasoning step.
- Evidence-Based Evaluation: In this step, the model extracts relevant evidence from the premise data (e.g., scientific studies) and systematically evaluates the claim by comparing its key elements with the extracted evidence. The process is guided by instruction prompts such as: *"1. Identify the relevant facts in the study. 2. Evaluate each piece of information in the claim against the facts."*.
- Relation Prediction: In the final step, the reasoning process concludes with a concise classification (e.g., SUPPORT or CONTRADICT) expressed in natural language. This prediction is based on the previously generated terms interpretations and evidence analysis, which are sequentially chained into the input prompt to guide the model's final decision.



Claim: 10-20% of people with severe mental disorder receive no treatment in low and middle income countries.

1. **10-20%**: This refers to a statistical range indicating that between one-tenth to one-fifth of individuals are affected
2. **Severe Mental Disorder**: This encompasses serious psychiatric conditions that can considerably impair an individual's cognitive and emotional functioning.
3. **Receive No Treatment**: This indicates that individuals in this category are not accessing any form of mental health care.
4. **Low and Middle-Income Countries**: This classification refers to nations that have lower economic resources and healthcare infrastructure compared to high-income countries.

Study:
Title: Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys.

CONTEXT ... RESULTS The prevalence of having any WMH-CIDI/DSM-IV disorder in the prior year varied widely, from 4.3% in Shanghai to 26.4% in the United States, with an interquartile range (IQR) of 9.1%-16.9%. Between 33.1% (Colombia) and 80.9% (Nigeria) of 12-month cases were mild (IQR, 40.2%-53.3%). Serious disorders were associated with substantial role disability. Although disorder severity was correlated with probability of treatment in almost all countries, 35.5% to 50.3% of serious cases in developed countries and 76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview. .... CONCLUSIONS Reallocation of treatment resources could substantially decrease the problem of unmet need for treatment of mental disorders among serious cases.

Supportive Evidence:
The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions.

Contradicting Evidence:
The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the problem.

Lack of Sufficient Evidence:
While there is detailed information about the treatment rates of severe mental disorders across different countries, the study does not provide a breakdown that specifically isolates the 10-20% statistic claimed. Therefore, while it does support the idea that a significant number of people do not receive treatment, it presents a broader and more alarming picture, particularly in low and middle-income countries.

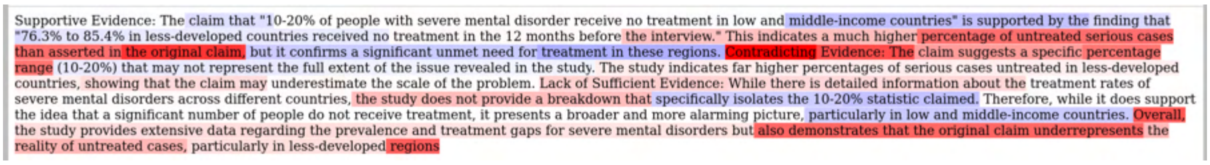
Conclusion: the study provides extensive data regarding the prevalence and treatment gaps for severe mental disorders but also demonstrates that the original claim underrepresents the reality of untreated cases, particularly in less-developed regions.

{"Relation": "Contradict"}

Figure 4: When prompting the LLMs with **CoENLI** framework, the process begins with *Semantic Grounding* and *Evidence-based Evaluation* steps. These steps help interpret key terms and assess each piece of claim against identified relevant data points. The highlighted words and phrases in the claim, study, and generated evaluation are intended to offer plausible insights involved in the claim verification process.

2.3. SHAP Values for Interpreting Word-Level Contributions

CoENLI enables LLMs to generate intermediate, evidence-based rationales and provide human-readable explanations of how the model processes claims and evidence. As illustrated in Figure 4, the *Evidence-Based Evaluation* step allows for broad reasoning, including identifying relevant evidence and evaluating both supportive and contradictory information. However, consolidating this evaluation into a final relation remains opaque, leaving interpretability gaps about which aspects of the evaluation contribute most to the final result. To address these limitations, we incorporate SHAP explanations using modules from [26][1], specifically developed to explain language models. By analyzing the Shapley values (SHAP) associated with the words in the input prompt, we can identify which features (words or phrases) generated in the intermediate step had the most significant influence on the final output of the generative model. Figure 5 illustrates feature relevancy based on SHAP values uisng a Mistral model [27] as explainer for the final relation result of the example from Figure 4. The saliency maps provide detailed insights into the model's decision-making process when balancing supportive and contradictory evidence to determine the relationship between the example claim and the study.



Supportive Evidence: The claim that "10-20% of people with severe mental disorder receive no treatment in low and middle-income countries" is supported by the finding that "76.3% to 85.4% in less-developed countries received no treatment in the 12 months before the interview." This indicates a much higher percentage of untreated serious cases than asserted in the original claim, but it confirms a significant unmet need for treatment in these regions. Contradicting Evidence: The claim suggests a specific percentage range (10-20%) that may not represent the full extent of the issue revealed in the study. The study indicates far higher percentages of serious cases untreated in less-developed countries, showing that the claim may underestimate the scale of the problem. Lack of Sufficient Evidence: While there is detailed information about the treatment rates of severe mental disorders across different countries, the study does not provide a breakdown that specifically isolates the 10-20% statistic claimed. Therefore, while it does support the idea that a significant number of people do not receive treatment, it presents a broader and more alarming picture, particularly in low and middle-income countries. Overall, the study provides extensive data regarding the prevalence and treatment gaps for severe mental disorders but also demonstrates that the original claim underrepresents the reality of untreated cases, particularly in less-developed regions

Figure 5: Words with positive SHAP values (highlighted in red) indicate phrases within the model-generated evaluation that significantly contribute to the Contradict classification, while words with negative SHAP values (highlighted in blue) indicate elements that reduce the likelihood of this classification.

3. Evaluation

Biomedical claim verification can be defined as logical relationship classification problem, where an NLI model determines whether a claim (C) logically follows from the premise evidence (P) provided in clinical trial or scientific study data. In our evaluation, we denote:

$$f(C,P) = \begin{cases} \text{Support} & \text{if } C \text{ logically follows} \\ & \text{ from } P; \\ \text{Contradict} & \text{otherwise} \end{cases} \qquad (1)$$

3.1. Datasets and Models

In order to assess the generalization capabilities of **CoENLI** across different LLMs, we evaluate their performance on two related benchmarks **NLI4CT** [4, 9] and **SciFact** [1, 28]. The claims from both datasets are written by human experts given clinical trial reports or scientific studies respectively. The **NLI4CT** challenges highlighted the difficulties of applying NLI models to validate claims related to clinical trial reports (CTRs). It requires a deep understanding of medical and scientific knowledge to interpret implicit data points beyond simple text matching. The premises in **SciFact** consist of evidence sentences extracted from the abstract of relevant studies. Wadden et al. [28] demonstrated the advantages of incorporating document-level premises compared to sentence-level premises for the **SciFact** challenge. Table 2 summarizes the number of instances for each relation class of the datasets applied in our evaluation.

In order to increase the trustworthiness of LLMs' outcomes, our system employs different lightweight open-source LLMs [29, 27, 30, 31]. The models applied are instruction-tuned [32] and compatible with the *FastLanguageModel* modules of unsloth.ai [33] for faster running on a single NVIDIA A100–80GB

[1] https://shap.readthedocs.io/en/latest/text_examples.html

Dataset	Support / Entailment	Contradict
NLI4CT (test set)	250	250
SciFact (dev set)	216	122

Table 2
The number of evaluation instances. **SciFact**'s test set withholds ground truth labels for leaderboard submissions [1], here we use its dev set as substitute.

GPU. Table 4 in Appendix A provides the version information about the applied models. In the evaluation, we compared their performance with two low-cost GPT models [34].

3.2. Results

To evaluate the reasoning capabilities of LLMs in a straightforward manner, We report prediction accuracy using F1 scores as detailed in Table 3.

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

The F1 scores are calculated using the scikit-learn library[2]. Our evaluation compares the performance of **CoENLI** against two baseline prompting methods:

- Simple prompt: A task-specific prompt template *"Return the logical relation between the provided claim and study: <Support> or <Contradict>."*. This represents a minimal and direct approach to the task with LLMs.
- zero-shot CoT: Building on the simple prompt, we introduce an additional instruction: *"Evaluate the relation step by step."* as proposed by Kojima et al. [22], prompting LLMs to deliver an intermediate reasoning process before responding the final relationship.

The comparison highlights the impact of task-adaptive **CoENLI** framework on the prediction accuracy of LLMs.

Model	NLI4CT				SciFact			
	Simple	CoT	**CoENLI**	**CoENLI***	Simple	CoT	**CoENLI**	**CoENLI***
GPT3.5	0.52 ± 0.01	0.53 ± 0.00	0.75 ± 0.01	0.82 ± 0.00	0.51 ± 0.03	0.76 ± 0.00	0.86 ± 0.00	0.88 ± 0.01
GPT4o-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	-
Llama3.1-8B	0.47 ± 0.00	0.54 ± 0.01	0.67 ± 0.02	0.80 ± 0.00	0.53 ± 0.02	0.80 ± 0.01	0.84 ± 0.05	0.90 ± 0.01
Gemma2-9B	0.63 ± 0.00	0.67 ± 0.03	0.75 ± 0.03	0.80 ± 0.00	0.57 ± 0.01	0.73 ± 0.00	0.86 ± 0.02	0.89 ± 0.01
Mistral-12B	0.55 ± 0.00	0.65 ± 0.01	0.75 ± 0.01	0.82 ± 0.00	0.65 ± 0.01	0.83 ± 0.00	0.87 ± 0.02	0.89 ± 0.00
Phi3-14B	0.62 ± 0.01	0.64 ± 0.00	0.75 ± 0.02	0.82 ± 0.00	0.76 ± 0.03	0.80 ± 0.01	0.88 ± 0.02	0.90 ± 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. **CoENLI*** represents the results of applying GPT4o-mini in the *Evidence-Based Evaluation* step. All the results demonstrate the high accuracy of the inference capability of GPT4o-mini model in the **CoENLI** framework in zero-shot setting.

While **CoENLI** demonstrates the enhanced performance of LLMs in the claim verification task compared to the baseline methods, a performance gap persists between GPT4o-mini and small-scale LLMs. Lightweight LLMs face challenges in delivering high-quality text analysis and often require fine-tuning with additional task-specific training examples [35]. Despite these limitations, smaller open-source models offer flexibility for SHAP explanations and can be optimized for specific tasks through the collection of training data during long-term development, increasing the controllability of the system. The choice between fine-tuning lightweight LLMs or incorporating GPT4o-mini's evaluations

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

into the **CoENLI** pipeline depends on the application's requirements and resource constraints. For our user study, we leverage GPT4o-mini's robust reasoning capabilities for the *Evidence-Based Evaluation* step. These outputs are then passed to small-scale models, which focus on generating final decisions and adding a second layer of interpretability using SHAP values. This hybrid approach combines the strengths of advanced and lightweight models to achieve both accuracy and transparency.

3.3. User Study

In constructing the datasets, Wadden et al. [1] reported a Cohen's kappa of 0.75 as inter-annotator agreement. Similarly, Jullien et al. [4] conducted a human evaluation of the **NLI4CT** task with three experts, achieving an average accuracy of 85% against the gold labels, the inter-annotator agreement with a Cohen's kappa of 0.83 . These results highlight the inherent variability in human judgement. This variability is often due to task uncertainty and differences in individual knowledge.

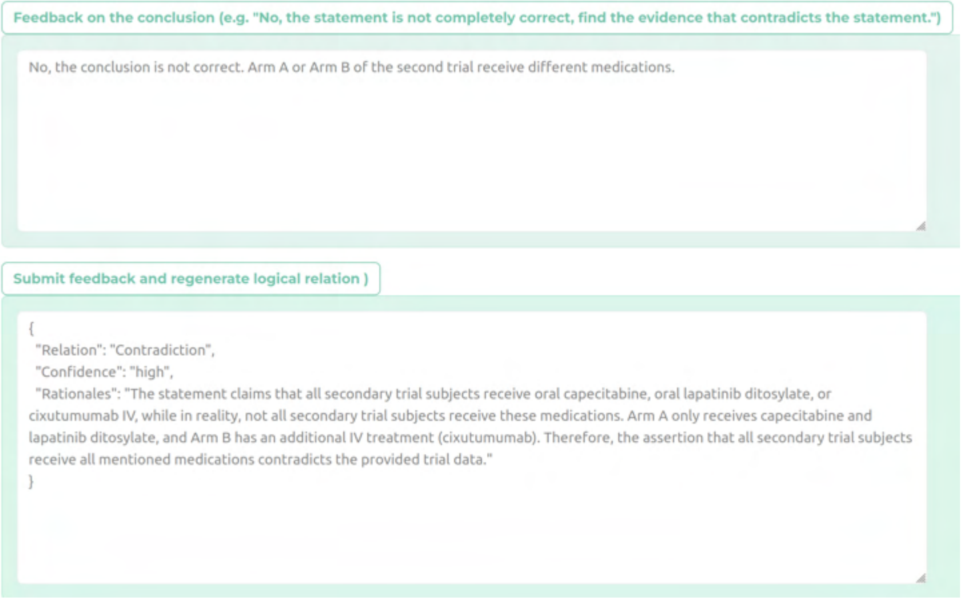
To evaluate the utility of the explainable claim verification system, we employ four medical students, each tasked with selecting a verdict for 20 claims from the **SciFact** test set (which lacks ground truth annotations) against five retrieved studies independently, resulting in 100 claim-study pairs in total. The user interface used for the study is shown in Figure 1. We observed an increase in inter-annotator agreement, with Cohen's kappa rising from 0.74 (without the Evidence-Based Explanation and SHAP values explanation) to 0.81. This improvement suggests that the system's enhanced interpretability fosters trust in the LLM outcomes and less modification of the initial results generated by the LLM, therefore promoting better alignment among users.

To assess the transparency and understandability of the model's reasoning process leading to the final relation classification, participants rated the model-generated rationales on a scale from 1 to 5. A score of 1 indicated that the reasoning process was confusing, with no clear connection between the rationales and the final classification, while a score of 5 indicated that the reasoning was fully transparent and easy to follow. All participants rated the reasoning transparency as a **4**, suggesting that the model's reasoning process is generally perceived as transparent. When asked what could be improved, participants provided feedback indicating the need for better quality in the intermediate reasoning steps generated by the LLMs. One participant noted, *"The NLI model sometimes overlooked the smallest details in the claim."*. This highlights how enhanced interpretability can help identify limitations in reasoning of LLM. As emphasized by [36], improving the functionality of these model-generated explanations is crucial for fostering user confidence in the system.

3.4. Feedback Loop Integration

Based on the results of the user study, which revealed that the model sometimes misses small details in claim verification, we integrate a feedback loop into the verification system. This enhancement directly addresses the identified gaps by allowing users to provide detailed feedback on specific errors or overlooked aspects of the model's reasoning. Through this mechanism as illustrated in Figure 6, users can highlight inconsistencies or missing details with concise guidance, enabling the system to regenerate both the relationship classification and the corresponding justifications. This iterative process not only helps to correct errors in real time, but also guides the model to refine its reasoning, ultimately improving its accuracy and reliability in verifying biomedical claims.

Furthermore, these feedback-driven rationales can be leveraged to fine-tune the model, enhancing its ability to recognize and account for nuanced details, particularly in logical consistency comparisons. This human-in-the-loop approach fosters greater user engagement, promotes collaborative verification, and ultimately strengthens the reliability of LLM-based biomedical claim verification systems.


Feedback on the conclusion (e.g. "No, the statement is not completely correct, find the evidence that contradicts the statement.")
Submit feedback and regenerate logical relation)

No, the conclusion is not correct. Arm A or Arm B of the second trial receive different medications.

```
{
  "Relation": "Contradiction",
  "Confidence": "high",
  "Rationales": "The statement claims that all secondary trial subjects receive oral capecitabine, oral lapatinib ditosylate, or
cixutumumab IV, while in reality, not all secondary trial subjects receive these medications. Arm A only receives capecitabine and
lapatinib ditosylate, and Arm B has an additional IV treatment (cixutumumab). Therefore, the assertion that all secondary trial subjects
receive all mentioned medications contradicts the provided trial data."
}
```

Figure 6: Consider the following verification result of one claim - *"All the primary trial participants do not receive any oral capecitabine, oral lapatinib ditosylate or cixutumumab IV, in conrast all the secondary trial subjects receive these."* from **NLI4CT**: *{"relation": "Entailment", "rationales": "The claim can be logically inferred from the clinical trial data points provided. The data clearly show that primary trial participants do not receive any oral capecitabine, oral lapatinib ditosylate, or cixutumumab IV, while all secondary trial subjects receive these medications in either Arm A or Arm B of the trial. This conclusion is supported by the specific interventions outlined for each group in the trial."}*. However, this result overlooks quantitative consistency—specifically the mismatch between "all the secondary trial subjects" in the claim and the different medications in "Arm A" and "Arm B" of the secondary trial subjects. This inconsistency highlights the difficulty of accurately aligning numerical or quantifier details within the reasoning process of LLMs. Hence, we integrate a feedback loop to empower users to actively refine the model's reasoning process.

4. Related Work

Biomedical claim verification falls into the broader task of Fact-Checking [37]. Automated claim verification is seen as a potential solution to enhance the speed and comprehensiveness of fact-checking in high-demand healthcare field [15, 38]. Additional datasets have been constructed and advanced machine learning methods to drive progress in automated biomedical claim verification system [1, 3, 2, 4]. However, in real-world scenarios, the verdict of claims is rarely either Support or Contradict, but often partially correct, contextually dependent, or misleading without additional explanation. Nakov et al. [39] argued that automated claim verification systems may aim to provide nuanced understanding rather than binary classifications. Li et al. [40] introduced a self-checker framework leveraging LLMs, which includes an evidence-seeker module to extract relevant evidence sentences for a given claim from retrieved passages. The framework allows human workers to validate the verdict prediction alongside the presented evidence, ensuring a more reliable verification process.

Chen and Shu [41] discussed the opportunities and challenges introduced by LLMs for automated claim verification with LLMs. While LLMs have demonstrated robust reasoning capabilities and human-readable explanations, they also pose significant threats through the generated misinformation, raising concerns about the trustworthiness of applying LLMs in claim verification tasks. Huang et al. [36] also emphasized that transparency in both the models and the underlying technologies is crucial for fostering trustworthiness and proposed principles spanning multiple dimensions to examine LLMs' trustworthiness. Through an extensive analysis of 16 LLMs across over 30 datasets based on these principles, they found that proprietary LLMs generally outperform open-source models in trustworthiness, largely due to their superior functional capabilities.

5. Conclusion

In this work, we present an explainable biomedical claim verification system that integrates iterative human-AI collaboration, evidence-based rationales, SHAP saliency explanations, and comparative insights from multiple LLMs. Our approach introduces the **CoENLI** framework to improve transparency, accountability and adaptability so that domain experts can better trust and use the results provided by LLMs. SHAP values further clarify how specific components of the model-generated rationales contribute to the final prediction, enhancing interpretability and user confidence. We explore the combination of advanced reasoning capabilities with the most advanced LLMs, such as GPT4o-mini and the open-source lightweight LLMs for interpretability, to achieve a balance between accuracy and transparency in the verification process. Additionally, our user study demonstrates the system's practical benefits, as indicated by an increase in inter-annotator agreement and feedback emphasizing the usability and trustworthiness of the model's reasoning process. Nevertheless, there are areas for refinement of the intermediate reasoning steps to address user concerns about overlooked details in claims.

In future work, we aim to integrate this explainable verification system as a component within a broader evidence synthesis framework to support human-AI collaboration in tasks such as combating misinformation in healthcare domain. Future work will also focus on refining intermediate reasoning quality, optimizing lightweight LLMs through task-specific fine-tuning to further enhance system performance, accessibility, and trustworthiness in biomedical claim verification and beyond.

Limitations

First, the reliance on GPT4o-mini in the *Evidence-Based Evaluation* step imposes computational resource demands that may limit accessibility in low-resource settings. Furthermore, the current framework may still struggle with claims requiring highly nuanced or specialized domain knowledge, where additional fine-tuning or inclusion of external expert input may be necessary. Small-scale LLMs, though flexible, exhibit performance limitations without additional task-specific training data. Finally, while SHAP values provide an interpretive layer, their effectiveness depends on the quality and granularity of the generated rationales.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg. We also appreciate the support by a grant from Accenture Labs. We also gratefully acknowledge the support provided by a grant from Accenture Labs[3].

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT[4] and DeepL[5] in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content and take full responsibility for the publication's content.

[3] https://iml.dfki.de/news/autoprompt-aims-to-improve-chatgpts-analysis-of-clinical-data/

[4] https://chatgpt.com/

[5] https://www.deepl.com/en/write

References

[1] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609. doi:10.18653/v1/2020.emnlp-main.609.

[2] M. Sarrouti, A. Ben Abacha, Y. Mrabet, D. Demner-Fushman, Evidence-based fact-checking of health-related claims, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3499–3512. URL: https://aclanthology.org/2021.findings-emnlp.297/. doi:10.18653/v1/2021.findings-emnlp.297.

[3] A. Saakyan, T. Chakrabarty, S. Muresan, Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic, arXiv preprint arXiv:2106.03794 (2021).

[4] M. Jullien, M. Valentino, H. Frost, P. O'regan, D. Landers, A. Freitas, SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2216–2226. URL: https://aclanthology.org/2023.semeval-1.307/. doi:10.18653/v1/2023.semeval-1.307.

[5] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[6] D. Wadden, K. Lo, L. L. Wang, A. Cohan, I. Beltagy, H. Hajishirzi, Multivers: Improving scientific claim verification with weak supervision and full-document context, arXiv e-prints (2021) arXiv–2112.

[7] H. Liu, A. Soroush, J. G. Nestor, E. Park, B. Idnay, Y. Fang, J. Pan, S. Liao, M. Bernard, Y. Peng, et al., Retrieval augmented scientific claim verification, JAMIA open 7 (2024) ooae021.

[8] A. Romanov, C. Shivade, Lessons from natural language inference in the clinical domain (2018). URL: http://arxiv.org/abs/1808.06752. arXiv:1808.06752.

[9] M. Jullien, M. Valentino, A. Freitas, SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1947–1962. URL: https://aclanthology.org/2024.semeval-1.271/. doi:10.18653/v1/2024.semeval-1.271.

[10] M. Jullien, M. Valentino, A. Freitas, Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials, arXiv preprint arXiv:2404.04963 (2024).

[11] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, arXiv preprint arXiv:2212.10403 (2022).

[12] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, H. Chen, Reasoning with language model prompting: A survey, arXiv preprint arXiv:2212.09597 (2022).

[13] Y. Xia, R. Wang, X. Liu, M. Li, T. Yu, X. Chen, J. McAuley, S. Li, Beyond chain-of-thought: A survey of chain-of-x paradigms for llms, 2024. URL: https://arxiv.org/abs/2404.15676. arXiv:2404.15676.

[14] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[15] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7740–7754.

[16] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al., Trustllm: Trustworthiness in large language models, arXiv preprint arXiv:2401.05561 (2024).

[17] D. Sonntag, M. Barz, T. Gouvea, A look under the hood of the Interactive Deep Learning Enterprise (No-IDLE), Technical Report, German Research Center for AI, 2024.

[18] D. Sonntag, T. Gouvea, M. Barz, A. Anagnostopoulou, S. Liang, S.-J. Bittner, F. Scheurer, The Interactive Deep Learning Enterprise (No-IDLE) meets ChatGPT, Technical Report, German Research Center for AI, 2024.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[20] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).

[21] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 22199–22213.

[23] Z. Yu, L. He, Z. Wu, X. Dai, J. Chen, Towards better chain-of-thought prompting strategies: A survey, arXiv e-prints (2023) arXiv–2310.

[24] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al., Least-to-most prompting enables complex reasoning in large language models, arXiv preprint arXiv:2205.10625 (2022).

[25] D. Lei, Y. Li, M. Hu, M. Wang, V. Yun, E. Ching, E. Kamal, Chain of natural language inference for reducing large language model ungrounded hallucinations, arXiv preprint arXiv:2310.03951 (2023).

[26] SHAP, shap.readthedocs.io. available online, https://shap.readthedocs.io/, 2024. (accessed in December 2024).

[27] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[28] D. Wadden, K. Lo, L. L. Wang, A. Cohan, I. Beltagy, H. Hajishirzi, MultiVerS: Improving scientific claim verification with weak supervision and full-document context, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 61–76. URL: https://aclanthology.org/2022.findings-naacl.6/. doi:10.18653/v1/2022.findings-naacl.6.

[29] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).

[30] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024).

[31] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

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

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

[35] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, B. Zhou, Enhancing chat language models by scaling high-quality instructional conversations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3029–3051. URL: https://aclanthology.org/2023.emnlp-main.183/. doi:10.18653/v1/2023.emnlp-main.183.

[36] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al., Position: Trustllm: Trustworthiness in large language models, in: International Conference on Machine Learning, PMLR, 2024, pp. 20166–20270.

[37] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, N. A. Smith (Eds.), Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508/. doi:10.3115/v1/W14-2508.

[38] G. Wang, K. Harwood, L. Chillrud, A. Ananthram, M. Subbiah, K. McKeown, Check-covid: Fact-checking covid-19 news claims with scientific evidence, arXiv preprint arXiv:2305.18265 (2023).

[39] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, arXiv preprint arXiv:2103.07769 (2021).

[40] M. Li, B. Peng, M. Galley, J. Gao, Z. Zhang, Self-checker: Plug-and-play modules for fact-checking with large language models, arXiv preprint arXiv:2305.14623 (2023).

[41] C. Chen, K. Shu, Combating misinformation in the age of llms: Opportunities and challenges, AI Magazine 45 (2024) 354–368.

A. Models

Model	Version	Context Window	Parameters
GPT3.5	gpt-3.5-turbo-0125	16K	175B
GPT4o-mini	gpt-4o-mini-2024-07-18	128K	?
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 4
List of low-cost GPT models and lightweight open-source LLMs used in our experiments, and 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.