Scientific Claim Verification with fine-tuned NLI Models

Keywords: Claim Verification, Deep Learning Models, Natural Language Inference, PubMed, SciFact dataset.

Abstract: This paper introduces the foundation for the third component of a pioneering open-source scientific question-answering system. The system is designed to provide referenced, automatically vetted, and verifiable answers in the scientific domain where hallucinations and misinformation are intolerable. This Verification Engine is based on models fine-tuned for the Natural Language Inference task using an additionally processed SciFact dataset. Our experiments, involving eight fine-tuned models based on RoBERTa Large, XLM RoBERTa Large, DeBERTa, and DeBERTa SQuAD, show promising results. Notably, the DeBERTa model fine-tuned on our dataset achieved the highest F1 score of 88%. Furthermore, evaluating our best model on the HealthVer dataset resulted in an F1 score of 48%, outperforming other models by more than 12%. Additionally, our model demonstrated superior performance with a 7% absolute increase in F1 score compared to the best-performing GPT-4 model on the same test set in a zero-shot regime. These findings suggest that our system can significantly enhance scientists' productivity while fostering trust in the use of generative language models in scientific environments.

1 INTRODUCTION

In the scientific field, the veracity of information is paramount, especially as large language models (LLMs) become increasingly integrated into research methodologies. The paper [ANONYMIZED] outlines a novel open-source initiative designed to mitigate the risks of inaccuracies, or "hallucinations", in answers generated by LLMs in the biomedical domain. Central to this initiative is a sophisticated architecture comprising three core components: an information retrieval system that utilizes both semantic and lexical search combination techniques to retrieve scientific papers from PubMed[1]; a Retrieval Augmented Generation (RAG) module with a generative model fine-tuned to produce referenced answers based on the retrieved scientific papers; and a verification engine tasked with cross-checking these generated answers against scientific papers to ensure accuracy and to identify potential hallucinations. This system aims to enhance the productivity of researchers by providing reliable information and also to instill trust in the use of generative language models within the scientific domains where misinformation can have serious repercussions.

According to (Guo et al., 2022), claim verification is a process performed after a claim is generated and evidence is retrieved. The authors also segment the claim verification process into two components: *verdict prediction*, where claims are assigned truthfulness labels, and *justification production*, where explanations for verdicts must be generated (Guo et al., 2022). Our system generates claims using RAG and we use the retrieved documents from the PubMed repository as evidence. In this paper, we are going to perform the verdict prediction task by transforming it into the natural language inference (NLI) task to predict one of the labels: support, contradict, and no_evidence. The development of a model specifically tailored for textual entailment or NLI represents a crucial element of the verification engine in [ANONYMIZED]. In this paper, we are going to present our effort on enhancing and fine-tuning different models to achieve accuracy in the task of scientific claims verification.

By fine-tuning state-of-the-art models for the NLI task in the scientific field, we plan to allow our model to detect the subtle differences in the statement between claims supported by evidence, those that contradict the evidence, and those for which evidence is lacking. This will not only build upon the foundational work presented in [ANONYMIZED] for claim verification but also introduce a novel methodology and models designed to provide claim verification by fine-tuning models for textual entailment tasks in the biomedical domain. The model fine-tuned for textual entailment will enhance the automated scientific claim verification process in [ANONYMIZED], complementing retrieval and generation steps by verifying generated text.

Our work's contributions are twofold: firstly, we fine-tuned different deep learning models on the Sci-

[1]https://pubmed.ncbi.nlm.nih.gov/

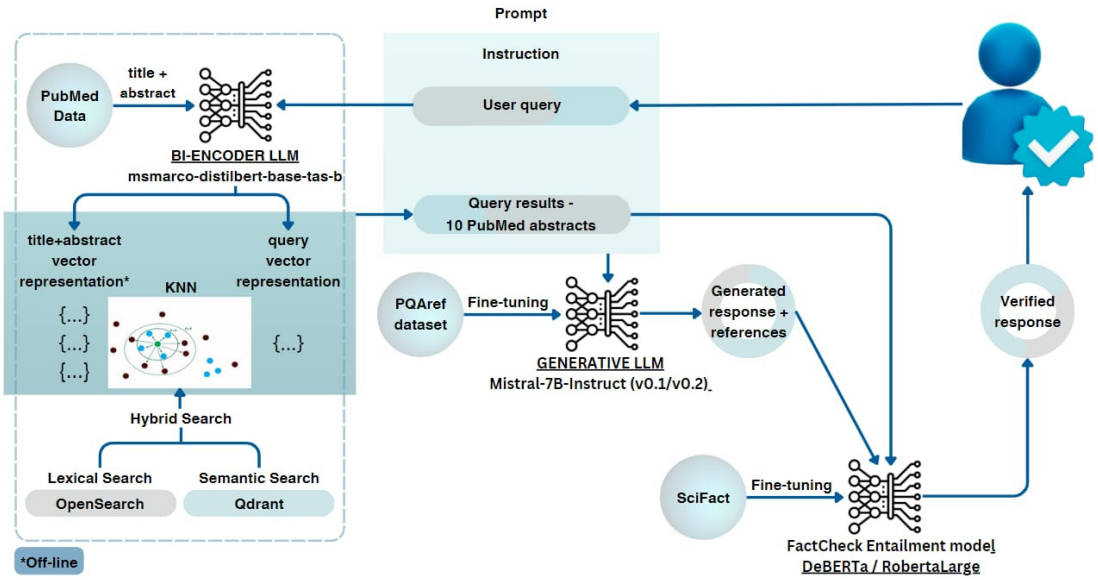

Figure 1: Architecture of our verification system.

Fact dataset to significantly improve textual entailment predictions in the biomedical domain, achieving state-of-the-art results. Secondly, we provide comparative results of these models fine-tuned with the SciFact dataset.

The paper is organized as follows: Section 2 provides an overview of our comprehensive system, detailing the RAG, and verification of claims produced. Section 3 presents current datasets and methods for the claim verification task. The transformation of the chosen dataset, as well as the models used for this task and their parameters, are presented in Section 4. Results are given in Section 5, with an accompanying error analysis in Section 6. Finally, Section 7 presents the conclusions drawn from our study and outlines directions for future research.

2 OVERALL SYSTEM DESIGN

Our verification system encompasses two main processes: Retrieval Augmented Generation (RAG) and the Verification Engine. The RAG process, consisting of two components, is based on a fine-tuned large language model (LLM) for referenced question-answering. In this setup, retrieved relevant abstracts from PubMed are provided to the LLM as input through a prompt. The output is an answer based on these PubMed abstracts, with each statement appropriately referenced, facilitating subsequent verification by the Verification Engine.

In this section, we describe the three main components of our system (see Figure 1), to offer clarity to the entire process which contains claim verification as its second and final verification step.

The Information Retrieval Component (IR) utilizes data from the PubMed database[2], which contains citations and biomedical literature from various sources. The IR system incorporates both sparse vectors (lexical index) and dense vectors (semantic index), facilitating lexical, semantic, and hybrid searches.

For lexical retrieval, based on BM25, we use OpenSearch[3] to create an index of PubMed articles by concatenating the titles and abstracts into a single indexed field. For semantic retrieval, based on dense vectors, we employ the Qdrant[4] vector database. To generate vector embeddings, we utilize a bi-encoder sentence transformer model pre-trained on the MS-Marco dataset[5], which had the highest performance on the Passage Retrieval Task at the time of indexing[6].

Hybrid search in our system combines the lexical and semantic IR components. To implement a hybrid search, we normalized the scores from these two IR

[2]https://pubmed.ncbi.nlm.nih.gov/download/

[3]https://opensearch.org/

[4]https://qdrant.tech/

[5]https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b

[6]https://www.sbert.net/docs/pretrained-models/msmarco-v3.html

methods to a scale ranging from 0 to 1. The scores from each search method are then multiplied by the respective importance weights. This approach allows for the identification of both direct matches and enhances the discovery of semantically related phrases and text segments, even when exact textual matches are absent. The selected documents from the IR Component are then passed to the Generative Component, which is responsible for generating the appropriate response.

The Generative Component receives the user query and 10 retrieved documents as its input. It consists of a generative model, currently Mistral-7B-Instruct-v0.2[7], which we additionally fine-tuned for the task of question-answering with references using the QLoRA methodology (Dettmers et al., 2023) and a dataset of randomly selected questions from the PubMedQA dataset (Jin et al., 2019). The output of the model is an answer to the user query which contains a reference for each of the claims generated based on the relevant articles.

The Verification Component is designed to verify the claims created by the RAG component. The specific type of verification and the models used for this process are described in detail in Section 4. Since the answer of the generative model can contain multiple claims/claim parts, our goal is to create a model for verifying individual claim parts, and how we combine them will be decided in the future.

3 RELATED WORK

The task of determining a claim's veracity is dubbed differently in the literature: from *verdict prediction* (Tan et al., 2023) or *veracity prediction* (Vladika et al., 2024) to *claim verification* (Wadden et al., 2020), to only name a few, and usually forms a part of a multi-component pipeline aiming to produce as factual results as possible (Tan et al., 2023). The task typically consists of assessing a claim and its corresponding evidence and categorizing it into one of three distinct labels: support/evidence, contradiction/refute, and no_evidence/not_enough_info. The process of verifying scientific claims can be conceptualized within the framework of natural language inference (NLI), treating the claim verification task as a multi-class classification task, a perspective that aligns with prior research (Thorne et al., 2018; Wadden et al., 2020).

Different techniques have been utilized and developed to improve claim verification, and recently, transformer models have achieved state-of-the-art performance in fact-checking in general, both in general and scientific domains (Tan et al., 2023). Input for these models usually consists of concatenated claim and evidence pairs from which they create representations for classifying relationships between the two (Tan et al., 2023). Including the entire context of the evidence (e.g. entire document) ensures minimum loss of information and better results during inference (Wadden et al., 2022).

Interestingly, both general-purpose and domain-specific large language models are used for the task of scientific claim verification (Vladika and Matthes, 2023). That general-purpose models are successfully used for this task is supported by the work of the creators of the SciFact dataset, who released the VeriSci model (Wadden et al., 2020). This model uses RoBERTa-large (Liu et al., 2019) model pre-trained on the FEVER dataset (Thorne et al., 2018) and fine-tuned on SciFact dataset as a component for label prediction, since it demonstrated the strongest performance compared to other tested scenarios. This model also showed better accuracy compared to SciBERT (Beltagy et al., 2019), BioMedRoBERTa (Gururangan et al., 2020), and RoBERTa-base when all trained only on the SciFact dataset.

While general-domain datasets for the task of claim verification existed since 2014 (Vlachos and Riedel, 2014), the first dataset for scientific claim verification, SciFact, appeared in 2020 (Wadden et al., 2020). The number of different scientific claim verification datasets continued to grow ever since, from those that collect claims from social media posts (Mohr et al., 2022), Wikipedia and Internet in general (Diggelmann et al., 2021; Sarrouti et al., 2021), different web portals (Kotonya and Toni, 2020; Vladika et al., 2024), science exam questions (Tan et al., 2023) or publications (Malaviya et al., 2023). However, SciFact is still one of the rare datasets that contain claims from research papers and is the most used dataset for building scientific claim verification systems to date (Vladika and Matthes, 2023).

The work of (Sarrouti et al., 2021) shows that the choice of the in-domain dataset for fine-tuning makes a significant difference. The authors conducted experiments on several baseline models: BERT (Devlin et al., 2019), SciBERT, BioBERT(Lee et al., 2019), and T5 (Raffel et al., 2020), trained and evaluated on the HealthVer dataset (Sarrouti et al., 2021), with T5 demonstrating superior performance over all other models. They also tested the BERT-base model fine-tuned on FEVER (Thorne et al., 2018), SciFact, PubHealth, and HealthVer datasets, and assessed its performance on the HealthVer test set. Their find-

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

ings suggest that despite the FEVER dataset's size advantage over SciFact and HealthVer, the model achieved superior F1 scores when trained on SciFact and HealthVer datasets. Since FEVER is based on Wikipedia sentences, this supports the notion that training on in-domain claims yields more substantial benefits for domain-specific claim verification tasks.

The positive effect of in-domain datasets is further confirmed by the work of (Tan et al., 2023), who performed in-domain fine-tuning first using Med-Fact and Gsci-Fact datasets (Tan et al., 2023) and then using SciFact, HealthVer, and CLIMATE-FEVER (Diggelmann et al., 2021) of BERT, DeBERTa (He et al., 2021), SciBERT, Longformer, and BioBERT. In comparison with only fine-tuning the models using SciFact, HealthVer, and CLIMATE-FEVER (Diggelmann et al., 2021), they managed to achieve an improved performance for most models, with DeBERTA performing best in almost all scenarios.

4 OUR MODEL FOR CLAIM VERIFICATION

Similarly to the prompt-generated answers to certain questions in generating long-form responses that contain multiple claims (Wei et al., 2024), the responses generated by our RAG module contain multiple claims supported by references to PubMed abstract as evidence. This structured approach ensures that each claim is supported by relevant scientific literature, enhancing the credibility and reliability of the generated responses.

The model's performance is evaluated through various metrics, including macro and weighted precision, recall, and F1-score, and also accuracy, which collectively provide a comprehensive understanding of its effectiveness in the task.

In our recent exploration of advanced natural language processing techniques, we opted for a fine-tuning process for 3 cutting-edge models: RoBERTa-large (Liu et al., 2019), XLM-RoBERTa Large (Conneau et al., 2019a), and DeBERTa Large (He et al., 2021), alongside DeBERTa SQuAD - a DeBERTa modelfine-tuned using the SQuAD dataset. Our primary objective was to enhance their performance on the task of textual entailment.

The selection of DeBERTa was motivated by its outstanding performance in claim verification experiments (Tan et al., 2023). RoBERTa was chosen for its modified attention mechanisms and the strongest performance on label prediction task (Wadden et al., 2020). XLM Roberta exhibited better performance on English data compared to RoBERTa (Conneau et al., 2019b). DeBERTa SQuAD was selected due to its training on question-answer pairs, specifically tailored for question answering tasks.

Fine-tuning was conducted using the SciFact (Wadden et al., 2020) dataset, which is specifically designed to support the development and evaluation of automated fact-checking systems focusing on scientific claims. Our objective in fine-tuning these models with the SciFact dataset was twofold: to improve their performance in discerning textual entailment within scientific texts and to compare their effectiveness in this specialized domain.

4.1 Dataset Transformation

The SciFact dataset comprises pairs of claims and evidence within the biomedical domain — a domain noted for its classification challenges even among human experts. We anticipate that it will serve as an optimal dataset for the fine-tuning of models aimed at claim verification within our system.

The SciFact dataset is structured into two separate files: *corpuses* and *claims*. The *corpuses* file contains the titles and abstracts of scientific articles, providing a rich source of evidence. The *claims* file includes various claims linked to these scientific articles via an identification number, facilitating the process of verifying the claims against the provided evidence. Initially, the claims file is divided into three parts: training, validation, and test. However, the publicly available version of the dataset does not include labels for the test partition. As a result, the 300 examples in the test partition were excluded from our analysis. To maximize the data available for our experiments, we combined the training and validation subsets, yielding a total of 1,711 examples. This combined dataset serves as the foundation for training, validating, and evaluating our claim verification models.

Initially, the titles and abstracts extracted from the corpus file underwent individual cleaning procedures. This process involved eliminating redundant spaces, excluding special characters present at the ends of the texts, removing surplus parentheses and brackets, and filtering out superfluous information contained within the abstracts, such as *(ABSTRACT TRUNCATED AT 250 WORDS)*. Recognizing the informative value inherent in article titles, we made a deliberate choice to concatenate them with their corresponding abstracts to construct a comprehensive response to the claims. Throughout this concatenation process, special attention was paid to titles lacking terminal punctuation. To maintain coherence and facilitate comprehension, we ensured that a concluding punctuation

mark was appended to such titles. This step was essential to mitigate potential ambiguity and prevent the model from misinterpreting the concatenated text in instances where unpunctuated titles were directly joined with the initial sentence of the abstract, creating potentially unrelated sentences.

Subsequently, using the identification number, we integrated information from both files, ensuring that each claim was matched with its corresponding concatenation of the title and abstract, along with one of three labels: *no evidence*, *support*, or *contradict*. The SciFact dataset was initially designed so that the same labels for an abstract were repeated if they were found in multiple sentences of that abstract. However, since we opted to consider the entire concatenated title and abstract as a single response to the claim, rather than treating each sentence separately, we performed a deduplication of these instances. This step ensured that redundant combinations were removed. Ultimately, this process yielded 1,213 unique claim+[title+abstract] combinations, which form the final dataset for our experiments. This comprehensive and deduplicated dataset provides a robust foundation for training and evaluating our claim verification models.

Descriptive statistics of the combined dataset formed in this manner revealed that approximately 36% of claim+[title+abstract] combinations are labeled as *no evidence*, about 42% are labeled as *support*, and around 22% are labeled as *contradict*. We divided the dataset into training, validation, and test subsets in a ratio of 80:10:10, ensuring that the proportion of each label is preserved across all subsets. This stratified division helps maintain the consistency of label distribution, which is crucial for accurately training and evaluating the model's performance.

4.2 Experiments and Training Parameters

Our study aims to perform a comparative analysis of various transformer models fine-tuned and evaluated using the additionally processed SciFact dataset.

Determining one of the three relationships (*no evidence*, *support*, *contradict*) in which a claim and an evidence can stand, we decided to consider as a Textual Entailment task. This task can be viewed as the (2) Sequence Classification task into 3 classes by appropriately structuring the input data. Therefore, the task of claim verification could be conceptualized as a multi-class classification task.

To address this task, transformer models were fine-tuned utilizing the concatenation of a claim (c) and a corresponding PubMed title+abstract concatenation, serving as evidence (e). Since the 4 models that we fine-tuned with the formatted SciFact dataset for the Textual Entailment task are rooted in either BERT or RoBERTa architecture, the inputs for them are:

$$[CLS]c[SEP]e[SEP]$$

for models with the basic architecture of the BERT model, and

$$< s > c < /s >< /s > e < /s >$$

for models with the basic architecture of the RoBERTa model.

The objective for our model is to accurately assign one of the following labels (l): *no_evidence*, *support*, or *contradict*. Based on the received input, the model's prediction is formalized as:

$$l\{c,e\} \in \{no\_evidence, support, contradict\}$$

This formulation allows for a systematic evaluation of the model's ability to discern and categorize the relationship between biomedical claims and their corresponding evidential support.

All trainings used ADAM optimizer (Kingma and Ba, 2014) with a learning rate value of 1e-5, weight decay of 0.01, and were conducted on a single DGX NVIDIA A100-40GB GPU using the PyTorch framework and Hugging Face Transformer library. The number of epochs for all models was initialized at 15, with an early stopping strategy implemented on the validation subset based on the F1 metric to determine the optimal model checkpoint. We appllied two distinct values for the early stopping hyperparameter for each of the 4 models (three and four), resulting in a total of eight fine-tuned models for evaluation. Model evaluation was conducted by exact prediction-label matching, employing standard performance metrics of F1 score, accuracy, precision, and recall.

5 RESULTS

Our models underwent a three-stage evaluation process. First, we assessed the models on a test subset of the transformed SciFact dataset (refer to Section 5.1) to evaluate their performance and identify the best model. Next, we tested the optimal model on an external dataset (Section 5.2). Finally, we compared the performance of our top model against GPT-4 models in Section 5.3.

Table 1: The results of eight fine-tuned models 80% of SciFact data, validated on 10% of SciFact data, and tested on remaining 10% of data

		RoBERTa L_{SF}				XLM RoBERTa L_{SF}				DeBERTa$_{SF}$				DeBERTa SQuAD$_{SF}$			
		NE*	S	C	wa	NE	S	C	wa	NE	S	C	wa	NE	S	C	wa
3	P	0.71	0.55	0.00	0.48	0.83	0.69	0.54	0.71	0.83	0.86	0.85	0.84	0.86	0.90	0.82	0.87
	R	0.73	0.82	0.00	0.61	0.89	0.67	0.52	0.71	0.86	0.84	0.81	0.84	0.86	0.88	0.85	0.87
	F1	0.72	0.66	0.00	0.53	0.86	0.68	0.53	0.71	0.84	0.85	0.83	0.84	0.86	0.89	0.84	**0.87**
	Acc	0.61				0.71				0.84				0.87			
4	P	0.85	0.75	0.67	0.77	0.75	0.76	0.71	0.74	0.88	0.90	0.88	0.89	0.82	0.91	0.88	0.87
	R	0.89	0.76	0.59	0.77	0.91	0.67	0.63	0.75	0.95	0.88	0.78	0.89	0.93	0.84	0.81	0.87
	F1	0.87	0.76	0.63	0.77	0.82	0.71	0.67	0.74	0.91	0.89	0.82	**0.88**	0.87	0.88	0.85	0.87
	Acc	0.77				0.75				0.89				0.87			

* NE: no_evidence, S: support, C: contradict, wa: weighted average, P: precision, R: recall, F1: F1 score
Acc: accuracy

Table 2: Results of the DeBERTa model fine-tuned on the 80% and 90% of the SciFact dataset end evaluated on the HealthVer test set.

	DeBERTa$_{SF-80}$				DeBERTa$_{SF-90}$			
	NE	S	C	wa	NE	S	C	wa
P	0.46	0.70	0.66	0.60	0.47	0.67	0.69	0.59
R	0.94	0.25	0.15	0.50	0.88	0.29	0.27	0.52
F1	0.62	0.37	0.24	**0.44**	0.61	0.40	0.39	**0.48**
Acc	0.50				0.52			

5.1 In-domain Evaluation

As indicated in Table 1, the best-performing model with an early stopping patience of 3 was DeBERTa SQuAD, achieving an F1-score of 0.87. In comparison, the DeBERTa model with an early stopping patience of 4 achieved the highest F1-score of 0.88.

The results also reveal that the *CONTRADICT* class poses the most significant challenge for the models, which is anticipated given that this class constitutes only 22% of the dataset, resulting in fewer examples for training and evaluation. This imbalance likely contributes to the models' difficulties in accurately predicting this class, highlighting the need for more targeted strategies to improve performance in underrepresented categories.

5.2 Out-of-domain Evaluation

To evaluate our best-performing model, DeBERTa fine-tuned with an early stopping patience parameter set at 4, on a dataset distinct from the one used for training and in-domain evaluation, we chose the HealthVer dataset. This dataset is designed for evidence-based fact-checking of health-related claims, allowing researchers to assess the validity of real-world claims by evaluating their truthfulness against scientific articles.

As can be seen in Table 2 (DeBERTa$_{SF-80}$), we obtained a weighted average F1 score of 0.44 and an accuracy of 0.50. Comparing these results to those reported by the authors in (Sarrouti et al., 2021), who fine-tuned a BERT-base model with SciFact and evaluated it on the HealthVer test set, we observe that they achieved the F1 score of 0.36 and an accuracy of 0.39. This demonstrates that our DeBERTa model fine-tuned on the transformed SciFact dataset improved upon these results.

In Section 5.1, we identified DeBERTa fine-tuned with 80% of the transformed SciFact dataset as the optimal model for Textual Entailment. Subsequently, we evaluated this model out-of-domain on the HealthVer dataset, observing superior performance compared to previous state-of-the-art (SOTA) models, with an absolute increase of 8% in F1 score. Given that the test subset, constituting 10% of the transformed SciFact dataset, had already been utilized for in-domain evaluation, we incorporated it into the

Table 3: Comparison of the DeBERTa$_{SF}$ model with GPT-4 models.

	DeBERTa$_{SF}$				GPT-4				GPT-4 Turbo				GPT-4o			
	NE	S	C	wa	NE	S	C	wa	NE	S	C	wa	NE	S	C	wa
P	0.88	0.90	0.88	0.89	0.85	0.77	0.84	0.82	0.93	0.81	0.65	0.82	0.72	0.91	0.74	0.80
R	0.95	0.88	0.78	0.89	0.80	0.94	0.59	0.81	0.64	0.92	0.81	0.80	0.89	0.80	0.63	0.80
F1	0.91	0.89	0.82	**0.88**	0.82	0.85	0.70	**0.81**	0.76	0.86	0.72	0.79	0.80	0.85	0.68	0.79
Acc	0.89				0.81				0.80				0.80			

training set. Subsequently, we retrained the DeBERTa model on 90% of the data from the transformed SciFact dataset. Upon evaluating the new model on the HealthVer dataset, we observed a further absolute improvement of 4% in the F1 metric (refer to Table 2, DeBERTa$_{SF-90}$ model). Furthermore, the exploration of augmenting the training dataset underscores the adaptability and robustness of our methodology.

5.3 Comparison with GPT-4 Models

We utilized the same test set as for our in-domain evaluation – 10% of our transformed SciFact, comprising 122 examples and encompassing three classes, to assess the performance of GPT-4, GPT-4 Turbo, and GPT-4o in zero-shot mode. The specific prompt employed for this testing was as follows:

> Critically asses whether the statement is supported, contradicted or there is no evidence for the statement in the given abstract. Output SUPPORT if the statement is supported by the abstract. Output CONTRADICT if statement is in contradiction with the abstract and output NO_EVIDENCE if there is no evidence for the statement in the abstract.

For all models, the temperature parameter was set to 0 to minimize randomness and generate the most deterministic outputs, while the max_tokens parameter was set to 350 to allow sufficient context generation. This approach enabled us to directly compare the performance of our fine-tuned transformer-based model with that of the GPT-4 series models in zero-shot regime under identical conditions.

In our experiments, we observed that our transformer-based model for claim verification outperformed GPT-4, GPT-4 Turbo, and GPT-4o (refer to Table 3). Specifically, our model demonstrated superior performance across various evaluation metrics, including both accuracy and F1-score. These results highlight the efficacy of our fine-tuning approach and the robustness of our model architecture in handling the complexities of claim verification tasks. The consistent outperformance of our model over the aforementioned state-of-the-art models underscores its potential for real-world applications and further establishes its credibility within the domain of automated fact-checking.

Additionally, our model is open-source, providing transparency and flexibility that are crucial for industries such as pharmaceuticals and biomedicine, where stringent process control is required. Unlike closed models, our open-source solution allows for comprehensive customization and verification, ensuring that the claim verification process adheres to the rigorous standards necessary in these fields.

6 ERROR ANALYSIS

An error analysis was undertaken to scrutinize misclassified claims within the in-domain evaluation subset of the transformed SciFact dataset, leveraging our top-performing model, DeBERTa$_{SF}$. As depicted in Figure 2, a total of 14 claim-abstract pairs were inaccurately classified, revealing the distribution of errors across different classes. While the model exhibits commendable performance in the NO_EVIDENCE and SUPPORT classes, it demonstrates a relatively higher misclassification rate in the CONTRADICT class, suggesting a focal point for potential enhancement.

The frequent misclassification of the SUPPORT class as the NO_EVIDENCE class primarily stems from the inclusion of numerical data in the evidence, encompassing various measures and variations. Additionally, the model's inability to recognize abbreviations of specific terms as equivalent contributes to this misclassification. Moreover, the presence of intricate details such as biological processes and chemical reactions may confound the model, especially when it lacks explicit fine-tuning to handle such nuanced information.

On the other hand, instances where genuine SUPPORT claims are incorrectly classified as CONTRADICT pose significant challenges within our dataset.

These misclassifications are often attributed to the semantic complexity inherent in clinical trial data, characterized by complex immunology terminology, detailed descriptions, and intricate comparisons. Furthermore, the model encounters difficulties in discerning the alignment between specific time frames provided in the evidence and categorizing the claim as contradictory, even when the evidence unequivocally supports the claim.

Notably, the CONTRADICT class exhibits the highest proportion of errors, with instances erroneously classified as SUPPORT in 4 examples and as NO_EVIDENCE in 2 examples out of a total of 27 CONTRADICT examples in the test set. In light of these observations, we will delve into the most problematic cases within this class.

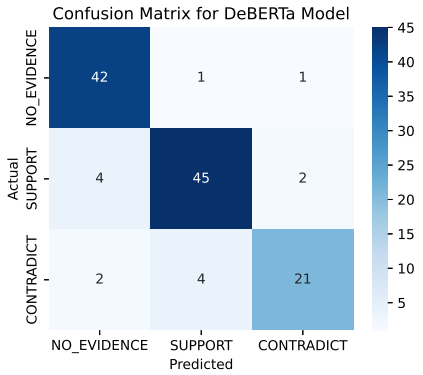

Figure 2: Confusion matrix for DeBERTa$_{SF}$ model.

Misclassifying claims from CONTRADICT to SUPPORT poses significant challenges, particularly in practical applications. These instances often exhibit surface-level similarities, marked by the presence of common keywords and phrases, without sufficiently interpreting the underlying semantic relationships. Consequently, the model may overlook clear contradictions in the evidence, relying instead on generalizations stemming from the presence of related terms. Furthermore, this misclassification is exacerbated by the semantic complexity inherent in the content of the evidence, along with the intricate analytical nuances embedded within scientific concepts. Additionally, challenges arise in comprehending the context surrounding complex biological processes, further complicating the accurate classification of claims.

For example, considering the claim "The genomic aberrations found in matasteses are very similar to those found in the primary tumor", several reasons contribute to the model misclassifying it as supporting the claim:

- The word "metastases" is incorrectly written as "matasteses" in the claim, but correctly spelled in the evidence, which could confuse the model.
- The claim implies a straightforward similarity between the genomic makeups of metastases and primary tumors. However, the evidence delves into the complex evolutionary process and genetic diversity of metastases, discussing timing, routes of dissemination, and genetic signatures of metastatic processes.
- The evidence describes metastases evolving in parallel, while the claim suggests a linear evolution, a critical distinction the model fails to recognize.

Misclassifying claims from CONTRADICT to NO_EVIDENCE may stem from inadequate representation of nuanced contradictions in scientific contexts within the training data. The absence of sufficient examples illustrating subtle contradictions may limit the model's ability to accurately discern and classify such instances. Moreover, the presence of semantic complexity, numerical data in experiments, variations in experiment timelines, and implicit contradictions further exacerbate the model's challenges in inference. Consequently, the model may struggle to make accurate classifications in scenarios where these factors interact, leading to instances of misclassification within our dataset.

For instance, considering the claim "The most prevalent adverse events to Semaglutide are cardiovascular.", several factors contribute to the misclassification:

- The evidence includes various numerical data points, such as medication dosages (e.g., 2.5 mg, 5 mg), percentages of change in hemoglobin A1c levels, body weight changes, and percentages of patients experiencing adverse events, which poses a challenge for the model.
- The contradiction is not explicitly stated, as the claim focuses specifically on cardiovascular adverse events as the most prevalent for Semaglutide. However, the evidence primarily discusses Semaglutide's efficacy in glycemic control, mentioning adverse events more broadly, with an emphasis on gastrointestinal events.
- The evidence contains extensive detail about the clinical trial's setup, participant demographics, dosage specifics, efficacy outcomes, and overall adverse events, overwhelming the NLI model, which needs to filter out irrelevant information to focus on the claim.

7 CONCLUSION AND FUTURE WORK

This paper outlines the development and evaluation of a Verification Engine as part of an open-source scientific question-answering system. By fine-tuning models for the Natural Language Inference task on a processed SciFact dataset, we aimed to provide referenced, automatically vetted, and verifiable answers.

Our in-domain evaluation identified the DeBERTa model, fine-tuned with 80% of the transformed SciFact dataset, as the optimal performer with an F1 score of 0.88. Testing this model on the HealthVer dataset, we achieved an F1 score of 0.44 and accuracy of 0.50, surpassing previous benchmarks and demonstrating significant improvements. Further experiments revealed that augmenting the training dataset to 90% led to an additional 4% increase in the F1 metric, emphasizing the robustness of our approach. Comparisons with GPT-4 models in a zero-shot regime showed our model's superior performance with a 7% absolute increase in F1 score to the best-performing GPT-4 plain model, highlighting its effectiveness in claim verification tasks.

Our model's open-source nature offers significant benefits for domains like pharmaceuticals and biomedicine, where rigorous process control is essential. Unlike closed models, our solution allows for transparency and customization, ensuring adherence to strict industry standards. Overall, our verification system enhances scientific productivity and establishes a reliable framework for automated fact-checking, crucial for maintaining the accuracy and integrity of scientific information.

Given, that the SciFact dataset contains challenging examples, we believe that the performance of the model tested on this dataset may be also underestimated compared to the real-world claims generated by a large language model. Nevertheless, we aim to further improve our methodology for claim verification in the future. We are planning to conduct research in the following directions:

- Generate a new dataset for claim verification - while SciFact is a decent dataset for biomedical claim verification using literature, we have noticed challenges in the dataset. Some claims may be short, and unclear without further context and this context is missing from the labels. We aim to collaborate with the industry and create a cleaner dataset, that overcomes these challenges.
- We understand that there are limitations of NLI, given even the most powerful neural architectures. Therefore, we aim to create a more comprehensive method, based on the combination of textual entailment task, text similarity, and chains of thoughts in fine-tuned large language models.

Enhancement in the dataset and a more comprehensive methodology will even further push the state-of-the-art in this domain and may contribute to the establishment of trust in generative search engines. Also, the area of claim verification is important as it may contribute to the adaptation of generative AI in the scientific domain, where the adaptation is currently limited due to the phenomenon of hallucinations, which makes verification of generated texts time-consuming and generated text unusable.

REFERENCES

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. (2021). CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. arXiv:2012.00614 [cs].

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020).

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kotonya, N. and Toni, F. (2020). Explainable Automated Fact-Checking for Public Health Claims. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Malaviya, C., Lee, S., Chen, S., Sieber, E., Yatskar, M., and Roth, D. (2023). Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Mohr, I., Wührl, A., and Klinger, R. (2022). CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(1).

Sarrouti, M., Ben Abacha, A., Mrabet, Y., and Demner-Fushman, D. (2021). Evidence-based Fact-Checking of Health-related Claims. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tan, N., Nguyen, T., Bensemann, J., Peng, A., Bao, Q., Chen, Y., Gahegan, M., and Witbrock, M. (2023). Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Vlachos, A. and Riedel, S. (2014). Fact Checking: Task definition and dataset construction. In Danescu-Niculescu-Mizil, C., Eisenstein, J., McKeown, K., and Smith, N. A., editors, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Vladika, J. and Matthes, F. (2023). Scientific Fact-Checking: A Survey of Resources and Approaches. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Vladika, J., Schneider, P., and Matthes, F. (2024). HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., and Hajishirzi, H. (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. Association for Computational Linguistics.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. V. (2024). Long-form factuality in large language models.