Leveraging fine-tuned Large Language Models with LoRA for Effective Claim, Claimer, and Claim Object Detection

Sotiris Kotitsas [†] [‡] **Panagiotis Kounoudis** [†] **Eleni Koutli** [†]
Haris Papageorgiou [†] [‡]
[†] Institute for Language and Speech Processing, Athena Research Center
[‡] Department of Informatics, National and Kapodistrian University of Athens
[sotiris.kotitsas,panagiotis.kounoudis,eleni.koutli,haris]@athenarc.gr
[skotitsas,xpapageor]@di.uoa.gr

Abstract

Misinformation and disinformation phenomena existed long before the advent of digital technologies. The exponential use of social media platforms, whose information feeds have created the conditions for many to many communication and instant amplification of the news has accelerated the diffusion of inaccurate and misleading information. As a result, the identification of claims have emerged as a pivotal technology for combating the influence of misinformation and disinformation within news media. Most existing work has concentrated on claim analysis at the sentence level, neglecting the crucial exploration of supplementary attributes such as the claimer and the claim object of the claim or confining it by limiting its scope to a predefined list of topics. Furthermore, previous research has been mostly centered around political debates, Wikipedia articles, and COVID-19 related content. By leveraging the advanced capabilities of Large Language Models (LLMs) in Natural Language Understanding (NLU) and text generation, we propose a novel architecture utilizing LLMs finetuned with LoRA to transform the claim, claimer and claim object detection task into a Question Answering (QA) setting. We evaluate our approach in a dataset of 867 scientific news articles of 3 domains (Health, Climate Change, Nutrition) (HCN), which are human annotated with the major claim, the claimer and the object of the major claim. We also evaluate our proposed model in the benchmark dataset of NEWSCLAIMS. Experimental and qualitative results showcase the effectiveness of the proposed approach. We make our dataset publicly available to encourage further research.

1 Introduction

In the information era where data are abundant and constantly flowing across the Internet, there is a high need for developing tools that evaluate the veracity of claims. Claim detection refers to the identification of potential claims in a given text and serves as an important component in both fact-checking and argumentation mining research. The process of evaluating a claim is a valuable and insightful procedure in various areas connected to fake news detection, misinformation, and disinformation prevention such as journalism (verification of sources), health (filtering of non-scientific claims), politics (discourse analysis) and marketing (detection of misleading claims). Additionally, there is a growing interest in targeting claims mentioned as "green" or "environmental", which mainly concern claims either related to statements of companies about eco-friendly products or to claims related to climate change.

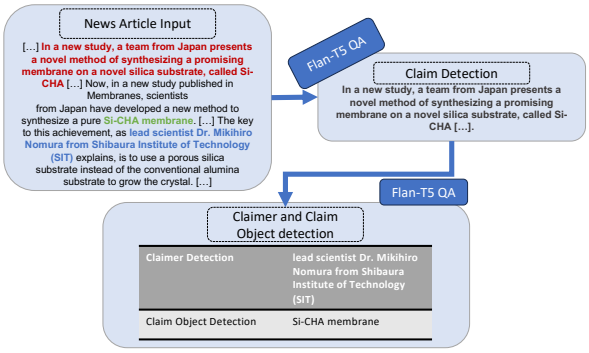

Figure 1: Example of a news article related to Health. The *major claim* is visible with red color, the *claimer* with blue color and the *claim object* with green color. Our task is to identify the *major claim* and extract related attributes, such as the *claimer* and the *claim object*.

Hassan et al. (2015) define a claim as "check-worthy" if the public is interested in determining its accuracy. In the argumentation mining theory, which regards *"the automatic identification and extraction of argument components and structure"* (Lawrence and Reed, 2019), claims are considered the fundamental components of an argument, and their identification is often closely linked to the specific context and topic (Levy et al., 2014, 2017; Gencheva et al., 2017; Aharoni et al., 2014). Although claim detection is involved in both argumen-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2540–2554

tation mining and fact-checking, a major difference is that argumentation mining does not necessarily require claims to be factual or verifiable (Reddy et al., 2021).

Claim detection also involves the identification of specific claim attributes, such as the claimer(s) and the object (topic) of the claim (Reddy et al., 2021; Li et al., 2022). A claimer can be an entity or the author of an article. Claimer detection is of major importance in order to assess both the credibility of the claim itself and the credibility of the source making it. Claim object identification has mainly been investigated as a preliminary step in the claim detection task, aimed at identifying claims in a specific domain, e.g., the COVID-19 pandemic (Reddy et al., 2021; Li et al., 2022).

Numerous research works have examined claim detection across different domains, mainly in political debates, Wikipedia articles, and most recently, COVID-19 related news articles (Hassan et al., 2017; Aharoni et al., 2014; Levy et al., 2014; Gencheva et al., 2017). Finally, Reddy et al. (2021) created NEWSCLAIMS, a COVID-19 related dataset, and evaluated various approaches for claim detection and attribute knowledge extraction (claimers, claim objects). However, NEWSCLAIMS uses predefined topics of COVID-19, making the task easier and limiting the models that adopt it, to specific domains. For example, in the claim *'One wild theory that has made its way around the web is that the virus came from space'*, along with the claim, the model is provided with the fact that the claim is related to *'origin of the virus'* and is basically tasked to identify the word *'space'*.

We propose a novel approach based on fine-tuned LLMs with LoRA, able to perform the task of *claim*, *claimer* and *claim object* detection by providing suitable prompts and without requiring a separate system for *claim detection* as in previous works. Our task is visible in Figure 1, where apart from the *claim* detection and additional attributes, we observe that we can perform all these tasks with a single model. Furthermore, unlike previous work, we also introduce a multi-domain dataset (HCN) containing general Health (not only COVID-19), Climate Change and Nutrition science related news articles. Each news article is annotated with the *major claim*, the *claimer*, and the *claim object*. We evaluate our model in HCN and NEWSCLAIMS demonstrating promising results in the three sub-tasks, showcasing the effectiveness of our methodology. We make the HCN dataset publicly available to encourage further research[1].

2 Related Work

The initial step in automated fact-checking is claim detection, where claims are identified for further verification (Guo et al., 2022). Typically, this detection process relies on the concept of checkworthiness. Previous works defined check-worthy claims as those that the general public would find interesting to know the truth about (Hassan et al., 2015). Konstantinovskiy et al. (2020) reframed claim detection as the determination of whether a claim presents an assertive statement about the world that can be fact-checked. They focused on whether the claim is verifiable using easily accessible evidence. Claims that rely on personal experiences or opinions are deemed uncheckable.

In the past, political debates were the main field of interest for claim detection. CLAIMBUSTER (Hassan et al., 2017), an end-to-end system for fact-checking, is trained on human-annotated sentences sourced from previous general election debates. CLAIMBUSTER incorporates "claim spotter", a component designed to determine the probability of a sentence containing claims that could be verified. Gencheva et al. (2017) introduced an openly available dataset derived from the 2016 US presidential and vice-presidential debates. They applied a context modeling approach on both their dataset and CLAIMBUSTER (Hassan et al., 2017) and state-of-the-art performance was achieved. Most recently, Jha et al. (2023) tried to detect check-worthy claims from Question-Hour debates of the Indian parliament, tweets posted by politicians, and Prime Minister statements.

Several studies have also focused on claims extracted from Wikipedia articles. Aharoni et al. (2014) presented a manually created argumentative dataset sourced from Wikipedia articles. The dataset comprises 2,683 argument elements that cover 33 controversial topics in claim-evidence pairs. Levy et al. (2014) utilized the above dataset to introduce the task of Context-Dependent Claim (CDC) Detection. They define CDC as *"a general, concise statement that directly supports or contests the given Topic"*.

Most recently, given the widespread impact of the COVID-19 pandemic, numerous works have emerged that specifically address claims related to COVID-19. Reddy et al. (2021) introduced NEWS-

[1] https://github.com/iNoBo/news_claim_analysis

CLAIMS, a manually created dataset for claim detection consisting of 143 COVID-19 related news articles with 889 annotated claims. Each claim has additional attributes, which are the claimer, the claim object, and the claim stance. Gangi Reddy et al. (2022) further used NEWSCLAIMS benchmark to evaluate on it a pre-trained Question-Answering (QA) system. Both works are thematically restricted, as they used a set of predefined topics regarding COVID-19. Similarly, COVID-19 Claim Radar (Li et al., 2022) is a system that integrates Claim Extraction and Knowledge Extraction to provide users with a structured and comprehensive understanding of claims associated with the COVID-19 pandemic.

CLEF CheckThat! is a shared task that began in 2018 and focuses on the automatic identification and verification of claims made on Twitter and political debates. Most recently, at CLEF CheckThat! 2022 (Nakov et al., 2022), where the focus was on tweets regarding COVID-19 and politics, transformer models such as BERT (Eyuboglu et al., 2022; Hüsünbeyi et al., 2022) and GPT-3 (Agrestia et al., 2022) achieved the best performance.

Lately, LLMs have also been leveraged for claim detection. Li et al. (2023) created the SELFCHECKER, a framework of plug-and-play LLMs modules for automatically fact-checking, which was evaluated on Fever (Thorne et al., 2018) and WiCE (Kamoi et al., 2023) datasets. Lu et al. (2023) also utilized LLMs and employed them to automatically generate claims for data augmentation.

Overall, up to now works on claim detection have focused on unstructured texts (e.g., political debates, tweets), while works on structured texts such as Wikipedia texts and news articles are domain specific, for example on the COVID-19 pandemic. Additionally, most of those approaches and existing automated claim detection systems prefer to define a factual and check-worthy claim, i.e. a segment of text containing measurable data, numbers or percentages, survey results, and relevant metrics. Those approaches may work well on unstructured text (tweets, speeches, etc.), but are weak when tackling structured text such as news articles. Furthermore, a non-specialist reader when confronted with texts presenting scientific data, research numerical results, etc. can benefit little from any verdict that a system gives on a particular claim containing such information.

Contrary to prior work, we locate the claim that summarises the main idea of a science-related news article, that is, the "major claim" of the text, since this is what a reader most easily identifies and is influenced by. The need for such an approach is high, as original findings from science publications may be distorted when reported to the public. This can lead to misinformation spread and, consequently, those altered versions of the original findings in the reporting may not be as accurate, possibly due to the different writing purposes between scientists and journalists (Li et al., 2017).

3 Datasets

In this section, we describe the datasets used in our methodology and experiments. We start with our HCN dataset and provide details on how we built it, and then move to the NEWSCLAIMS dataset, putting emphasis on its different sub-tasks and how these correlate with our task.

3.1 HCN Dataset

3.1.1 Motivation

Since the COVID-19 pandemic began, a great deal of information has circulated online in news articles, resulting in an infodemic. As a result a lot of research has been done since then in claim detection, specifically targeted to articles related only to COVID-19 (as mentioned in Section 2). On the other hand, our proposed dataset not only contains news articles in COVID-19 but also in General Health, Nutrition and Climate Change domains. Regarding the latter, there exists the EU initiative on green claims, which focuses primarily on press releases from businesses on their products and whether or not they are environmentally beneficial. Stammbach et al. (2023) and Diggelmann et al. (2020) focus only on the claim detection task and according to our knowledge there are no other datasets that focus on claim, claimer and claim object detection in the Climate domain. Statistics of HCN are presented in Table 1[2].

3.1.2 Annotation Guidelines

In HCN we define Major Claim, Claimer(s) and Claim Object(s), in the context of a scientific news

[2]For collecting the news articles, we utilized RSS feeds of news websites to obtain 867 scientific-related news articles covering the domains Health, Nutrition and Climate change and also covering a wide and diverse range of news outlets, which also encompass different journalists. Since we utilize LLMs and their tokenization is based on BPEs, no heavy text preprocessing is needed (lemmatization, stop-word removal, stemming etc.). Furthermore, the news articles stem from RSS feeds which require no text processing.

Statistics	Value
Total News Articles	867
Major Claims	867
Reported Claimers	1369
Claim Objects	1213

Table 1: Statistics of the HCN dataset. Each article has a major claim, possibly multiple claim objects associated with the claim and possibly multiple claimers.

article as follows[3]:

Major Claim: We consider as major claim, an argumentative sentence that includes the main point the claimer(s) want(s) to convey or to report, which presumably summarizes the study findings related to the news article.

Claimer: We define a *Reported* claimer as either a Person or an Organization that asserts a major claim within the text. We do not consider as claimers, plain artifact snippets such as "according to the study", "a report said", since they do not provide sufficient information about the entity that makes the claim. However, we annotate artifact snippets tied to named entities (e.g. *a study conducted by University of XX*). When the news article reports on a study, we annotate every claimer who is related to it (co-authored) and we prioritize its most informative occurrence (e.g. *'XX senior author of the study and researcher at the University of XX'*). When none of the above conditions is met, then the claimer of the article is considered to be the *Journalist.* Finally, note that the claimers annotated can be present or not in the major claim.

Claim Object: One or more snippets of text that most aptly describe(s) the theme (topic) of the claim identified. Note that the claim objects may appear elsewhere in the news article and are not necessarily included in the major claim sentence. However, we prioritize annotating claim objects present in the major claim.

3.1.3 Annotators

The 2 annotators used are linguists, post-graduate students in language technology, one woman 25 years old and one man 29 years old [4]. Their role was to annotate the 867 news articles of HCN [5]. Their work comprised of initially annotating the same sample of 100 news articles so we can calculate *inter-annotator agreement* (IAA) scores and ensure that the annotation guidelines are clear. After substantial agreement was achieved in the sample of news articles, the rest of the dataset was evenly split amongst them. The 2 annotators did not collaborate at any point during the annotation process, to avoid influencing each others annotations.

3.1.4 Inter-Annotator Agreement

For the inter-annotator agreement (IAA) we calculate COHEN'S KAPPA scores. COHEN'S KAPPA is a statistical measure used to calculate IAA, particularly in tasks involving categorical annotations. It considers the agreement between two or more annotators, accounting for the possibility of agreement occurring by chance. Table 2 presents the COHEN'S KAPPA score for each attribute of the HCN dataset. We observe substantial agreement for all 3 of the attributes annotated, which is above the agreement expected by chance.

Attribute	COHEN'S KAPPA
Major Claim	0.84
Claimer	0.82
Claim Object	0.73

Table 2: IAA scores for the HCN dataset. The scores are calculated with COHEN'S KAPPA.

3.2 NEWSCLAIMS

In Reddy et al. (2021), NEWSCLAIMS extends claim detection by extracting additional background attributes related to the claim, such as claim objects and claimers. The evaluation of claim detection in NEWSCLAIMS focuses on an emerging real-world scenario, specifically claims related to various aspects of COVID-19[6].

Claim Sentence detection: The *claim sentence* detection sub-task in NEWSCLAIMS is to identify sentences that contain claims related to predefined aspects of COVID-19. They utilize the CLAIMBUSTER model to identify candidate claim sentences and then filter them according to the predefined aspects using Natural Language Inference (NLI) methods. We differentiate from the above-mentioned approach, since our main focus is to

[3] Refer to Appendix section A.6 for annotation examples regarding the major claim, claimers and claim objects of HCN.

[4] They are employees in the program that conducted the research. Refer to the Acknowledgement section.

[5] The 2 annotators used the Inception (Klie et al., 2018) tool for annotating the news articles. Furthermore, the same tool was used for calculating IAA scores.

[6] Origin, transmission, cure, and protection from the virus.

identify the major claims in news articles, without limiting the model to predefined scopes of the claims.

Claim Object detection: A *claim object* refers to what is being claimed, i.e. the topic of the claim. This sub-task of NEWSCLAIMS is somewhat similar to our sub-task of claim object detection. The fundamental difference is that in NEWSCLAIMS the models take as input not just the claim but the knowledge that the claim is related to a predefined COVID-19 aspect. We on the other hand, have news articles in two extra domains (Climate Change, Nutrition) and no predefined aspects, making the task inherently more challenging.

Claimer detection: In the *claimer* detection subtask of NEWSCLAIMS we try to identify the source of the claim. The source of the claim can be *reported*, meaning a person, an organization, an artifact etc. or it can be the *Journalist* i.e. the author of the news article. The fundamental difference with HCN is that HCN only contains entities as claimers (persons or organizations).

4 Methodology

Generative AI is undergoing impressive growth with LLMs leading the way. LLMs are intricate models consisting of billions of parameters, trained on extensive collections of text and have demonstrated exceptional effectiveness across a broad spectrum of text-related assignments. It is worth mentioning that these models still remain language models, trained on the text completion task (i.e. predict the next token), enabling them to learn and generalize from extensive and diverse training data. These models however must be fine-tuned in specific tasks in order to be effective and to better adapt to specific domains or types of text that were not well represented in their original training data. To that end, we employ *instruction fine-tuning* to tune the model using examples of the target task. Instruction fine-tuning involves utilizing a collection of labeled examples represented as *prompt-response* pairs to enhance the pre-trained model's ability to accurately predict the response based on the given prompt.

4.1 Instruction Finetuning with LoRA

Conventional fine-tuning of LLMs is not efficient because it is computationally expensive and resource-intensive, since it requires to update the parameters of the original model. As LLMs scale up, fine-tuning and storing all the parameters becomes prohibitively costly and eventually becomes practically infeasible (Ding et al., 2023). However, there are several parameter-efficient alternatives to conventional fine-tuning, such as prompt-tuning, adapters and LoRA for example. In this work, we utilize LoRA (Hu et al., 2021). LoRA freezes the pre-trained model weights and introduces trainable rank decomposition matrices into every layer of the Transformer architecture. This approach significantly decreases the number of trainable parameters for downstream tasks. The foundation model of the proposed work will be FLAN-T5 base (Chung et al., 2022). The selection of FLAN-T5 base was based on the good performance to parameter ratio, requiring considerable less computational resources to fine-tune and infer. Furthermore, the proposed methodology is agnostic to the LLMs selected.

In order to fine-tune FLAN-T5, our datasets must undergo a transformation process wherein the instances are converted to *instruction, answer* pairs. This transformation involves structuring the data such that each instance consists of an instruction and the corresponding desired answer (output). Since we aim at identifying claim sentences along with attributes (claimer and the claim object of the claim), we must create three separate instructions, which are illustrated in Table 7 in Section A.2. We observe that in the *Claimer detection* sub-task we instruct FLAN-T5 to answer *No claimer found* in case the claim does not have a claimer. This corresponds with NEWSCLAIMS where the author of the news article is considered as the claimer, when a claim does not have one. However, based on the claimer instruction in Table 7, we alleviate the need to fine-tune a threshold to decide whether the claim has a claimer or not, as was done in previous work (Reddy et al., 2021; Gangi Reddy et al., 2022).

Finally, it is worth mentioning that in the HCN dataset we always have one major claim, but we might have more than one claimers and claim objects as it is shown in Table 1. As a result, for the news articles that have more than one claimer (or claim object), we create as many instances as the claimers (or claim objects). Presumably the model will be able to learn, based on the context. For example, if the article has more than one claimers, when prompted at inference time, the model will answer *"researcher X and researcher Y"*. Examples of this behavior are presented in Section A.4.

4.2 Inference

After fine-tuning our foundation model FLAN-T5 with LoRA we are able to provide prompts to our model to generate answers. Regarding the sub-tasks of *Claimer* and *Claim object* detection, the inference of the model is straightforward. We treat these two sub-tasks as a QA problem. The input to the model is the news article along with the same instructions presented in Table 7 for the claimer and claim object. We also provide the claim, since different claims have different claimers and claim objects. Example inputs are presented in Section A.5.

As already mentioned, the news articles in the HCN dataset always have one major claim. However, this is not the case with NEWSCLAIMS, since each news article has multiple claims associated with predefined aspects of COVID-19. To be able to evaluate our proposed approach in NEWSCLAIMS, FLAN-T5 must be able to classify multiple sentences as relevant claims. Motivated by Reppert et al. (2023); Salazar et al. (2019) we try to generate probability scores for the text generated by FLAN-T5. In Salazar et al. (2019), the authors generated pseudo-log-likelihood scores (PLLs) by masking tokens one by one. We on the other hand, provide our model a set of possible answers (choices) and at each state of the text generation process, we can gather the log-likelihood of the token generated. As a result, by summing all the likelihood scores of the generated tokens each choice contains, we can calculate probability scores for each choice. We normalize the calculated scores for the possible answers. For the *claim detection* task, the choices provided to the model are the sentences of the news article, which we can now rank according to their calculated score. Sentences sharing a lot of tokens will be ranked high, resulting in similar claims. As a result, we perform the aforementioned procedure as many times as the claims we want to identify, removing the previous extracted major claim from the text of the news article.

5 Experiments

We investigate the effectiveness of our proposed approach in the two datasets, namely HCN and NEWSCLAIMS. In the Appendix section A.3 we outline the details of our implementation, the hyperparameter tuning performed in the HCN dataset and the optimal sets of parameters. In the next sections, we discuss the evaluation setting, the baselines used in the comparative analysis and the results of our experiments in the three sub-tasks outlined in this paper.

5.1 Baseline Models

As outlined in this paper, our main method is a foundation model FLAN-T5 fine-tuned using LoRA in the HCN dataset (FLAN-T5-LORA-HCN). To evaluate the benefits of fine-tuning in a specific task, we also provide results with a base FLAN-T5 (FLAN-T5-BASE-HCN). Regarding the NEWSCLAIMS dataset, the authors offer a small set of news articles (18 in total) for fine-tuning and a test set for evaluation. We use the abovementioned models in a zero-shot setting evaluating them in the test set of NEWSCLAIMS (denoted as FLAN-T5-BASE-ZERO-SHOT and FLAN-T5-LORA-HCN-ZERO-SHOT). We do not expect our zero-shot setting models to perform well in NEWSCLAIMS, since the sub-tasks between the two datasets have differences (as mentioned in Section 3). Subsequently, we provide evaluation results with two additional variants, namely FLAN-T5-LORA-HCN-NC and FLAN-T5-LORA-NC[7].

To compare our proposed models in NEWSCLAIMS we utilize the reported numbers[8] in Reddy et al. (2021) and Gangi Reddy et al. (2022). Reddy et al. (2021) utilized CLAIMBUSTER for *claim detection* and then filtered the claims according to predefined aspects of COVID-19 using NLI. For *claimer detection* they consider a Semantic role labeling (SRL) baseline and a BERT model trained in existing datasets (POLNEAR) (Newell et al., 2018), which uses a threshold to determine if the claim is by the *Journalist*. For *claim object detection* they employ GPT-3 and T5 in zero-shot prompting, few-shot prompting for in-context learning (Brown et al., 2020) and prompt-based fine-tuning (Gao et al., 2021). Gangi Reddy et al. (2022) proposed a framework utilizing a zero-shot BERT QA model pretrained in SQUAD (Rajpurkar et al., 2018) and Natural Questions (NQ) (Kwiatkowski et al., 2019)[9]. Following, Reddy et al. (2021), they also use predefined COVID-19 aspects and CLAIMBUSTER for claim detection, although they filter

[7]Where FLAN-T5-LORA-HCN-NC is the pre-trained FLAN-T5 in HCN, further fine-tuned in NEWSCLAIMS and FLAN-T5-LORA-NC is a FLAN-T5 base model fine-tuned with LoRA in NEWSCLAIMS.

[8]No available implementations exist for the models described in these two papers. However, the authors report updated numbers in their Github: Link

[9]Results from this paper will be indicated with QA or the postfix QA when necessary.

the claims utilizing their QA model. The *claim object detection* is performed through the QA model, however again utilizing the predefined aspects to extract a claim object related to said aspect. Finally, the *claimer detection* is solely performed through the QA model, where the answer is again thresholded.

On the contrary, our proposed framework is a single fine-tuned model capable of performing all three sub-tasks, by providing different instructions. We are not restricted from predefined aspects at any stage, meaning we can apply our model in different domains as seen in the HCN dataset. Additionally, for the identification of the claimer, we do not need to tune a threshold to decide whether the journalist or an entity mentioned in the news article makes the claim, since it is encoded in our model through the claimer instruction as already mentioned in Section 3.

5.2 Evaluation Setting

To evaluate the proposed approach and its variants we use the evaluation scripts provided from NEWSCLAIMS[10] and calculate token-wise F1 (denoted as F1*) scores for the *claimer* and *claim object* detection tasks. Furthermore, for the *claimer* identification and following previous work (Reddy et al., 2021; Gangi Reddy et al., 2022), we also calculate F1-IN-SENTENCE and F1-OUT-SENTENCE[11] for the NEWSCLAIMS dataset. The evaluation of *claim detection* differs in the two datasets. In the HCN dataset, we always have one claim, which is the major claim. As a result, we generate the claim using the claim instruction in Table 7 and calculate F1*. For the NEWSCLAIMS dataset, each news article contains multiple claims. As mentioned, we cannot generate multiple answers, hence the reason of creating the ranking mechanism described in Section 4.2. We calculate F1 when considering the TOP-K answers with k being equal to the number of claims each NEWSCLAIMS article has, to ensure fair comparison.

5.3 Results

5.3.1 Claim Sentence detection

Claim detection results for the HCN and NEWSCLAIMS datasets are presented in Tables 3 and 4 respectively. In the HCN dataset, we observe that the fine-tuned variant of FLAN-T5 using LoRA clearly outperforms the base FLAN-T5. This indicates that fine-tuning for specific tasks greatly improve the performance, even when using simple LLMs, also showcasing the efficacy of the HCN dataset. Regarding the NEWSCLAIMS dataset, all the variants of FLAN-T5 outperform CLAIMBUSTER, in F1 scores. The best recall is achieved however through CLAIMBUSTER, since it has inherently high recall and is trained in detecting *check-worthy* claims, with factual information. As a result, Gangi Reddy et al. (2022) use their QA model to filter claims related to the predefined COVID-19 aspects, outperforming all the other models in that comparison. Our models, perform fairly well with the proposed scoring mechanism, without the knowledge of predefined aspects. However, there is room for improvement, by exploiting larger LLMs.

5.3.2 Claim Object detection

Regarding *claim object detection* results are presented in Table 3 for HCN and Table 4 for NEWSCLAIMS. Again in the HCN dataset FLAN-T5-LORA-HCN outperforms by a large margin FLAN-T5-BASE-HCN. The inference procedure of these two models is generative. Correspondingly, their performance is in reality higher, if we account for synonyms and contextually similar answers with the true labels. Examples are presented in the Appendix A.4. In NEWSCLAIMS as expected our zero-shot models do not perform well. However, FLAN-T5-LORA-HCN-ZERO-SHOT outperforms GPT-3 and T5 (zero-shot), indicating the pre-training in HCN helps in this sub-task. Our fine-tuned FLAN-T5 variants (postfix NC) outperform all the models in this comparison with the pretrained variant in HCN being the best again, even though the QA model during inference takes as input the fact that the claim is related to *origin of the virus* for example and asked to identify the claim object. Fine-tuning even with low number of instances (18 news articles) is clearly beneficial for the sub-task of the *claim object* detection.

5.3.3 Claimer detection

Following the trend with the previous sub-tasks, FLAN-T5-LORA-HCN outperforms the base model by a considerable margin. In the NEWSCLAIMS dataset the settings of the task are similar to the HCN dataset, where the predefined aspects of COVID-19 are not utilized. However, the claimers annotated in these two datasets differ. As men-

[10] Github Link:Link

[11] Where F1-IN-SENTENCE considers the instances where the reported claimer resides inside the claim and F1-OUT-SENTENCE, where the claimer does not exist inside the claim.

Model	Claim	Claimer	Claim Object
	F1*	F1*	F1*
FLAN-T5-LORA-HCN	**72.91**	**76.59**	**74.71**
FLAN-T5-BASE-HCN	56.16	3.90	21.32

Table 3: Token-wise F1 (F1*) for the HCN dataset in Claim, Claimer and Claim object detection tasks. Numbers are in % and calculated in the test set.

Model	Type	Claim			Claimer					Claim Object
		F1	P	R	F1*	Reported	Journalist	F1-IN-SENTENCE	F1-OUT-SENTENCE	F1*
FLAN-T5-BASE-ZERO-SHOT	Zero-shot	30.98	30.92	31.00	20.10	2.40	44.90	2.30	2.40	5.70
FLAN-T5-LORA-HCN-ZERO-SHOT	Zero-shot	31.87	31.81	31.93	22.17	17.43	28.80	19.65	13.65	15.38
FLAN-T5-LORA-HCN-NC	Fine-tuned	32.76	**32.70**	32.82	50.20	44.57	58.00	54.00	28.48	**59.15**
FLAN-T5-LORA-NC	Fine-tuned	30.59	30.53	30.65	**51.50**	**46.85**	58.13	**60.27**	23.93	57.40
CLAIMBUSTER	-	22.60	13.00	**86.50**	-	-	-	-	-	-
CLAIMBUSTER + QA	-	**36.00**	30.70	43.40	-	-	-	-	-	-
SRL	Fine-tuned	-	-	-	41.70	23.50	**67.20**	35.80	2.40	-
POLNEAR	Fine-tuned	-	-	-	42.30	25.50	65.90	38.90	2.70	-
QA	Zero-shot	-	-	-	50.10	39.80	64.40	46.20	**29.00**	57.00
GPT-3	Zero-shot	-	-	-	-	-	-	-	-	15.20
T5	Zero-shot	-	-	-	-	-	-	-	-	11.40
GPT-3	Few-shot	-	-	-	-	-	-	-	-	51.90
T5	Fine-tuned	-	-	-	-	-	-	-	-	51.60

Table 4: Token-wise F1 (F1*), F1, P (Precision) and R (Recall) for the NEWSCLAIMS dataset in Claim, Claimer and Claim object detection tasks. For the claim detection task we calculate regular F1 and for the claimer and claim object detection tasks we calculate F1*. Numbers are in % and calculated in the test set. The token "-" is used to showcase that the method is not applicable to the setting.

tioned in Section 3.1.2, we only annotate entities as claimers and not artifacts. This is not the case with NEWSCLAIMS, since this dataset also contains artifacts as claimers (e.g. "study", "researchers" etc.). We observe in Table 4, that our model FLAN-T5-LORA-NC outperforms all the other models in overall F1* scores, with FLAN-T5-LORA-HCN-NC and QA performing nearly as well. FLAN-T5-LORA-HCN-NC does not perform the best in this sub-task, since it was fine-tuned in news articles only containing entities as claimers and the low number of instances containing artifacts as claimers in NEWSCLAIMS, are presumably not enough. The SRL model obtains the best results in the *Journalist* setting. The SRL baseline works in sentence-level utilizing predicates and cannot extract claimers outside of the claim sentence. As a result, claim sentences with no predicates or explicit claimer mention will be classified as *Journalist*. Gangi Reddy et al. (2022) also showed that claim sentences made from the author of the news article, are sentences from a first-person point of view, containing no predicates, further validating the performance of SRL in the *Journalist* setting. Finally, both SRL and POLNEAR are sentence-level baselines indicating their poor performance in F1-OUT-SENTENCE in Table 4. FLAN-T5-LORA-NC outperforms the other models in F1-IN-SENTENCE with FLAN-T5-LORA-HCN-NC performing similar as well and both of them having a significant performance gap from the other models. QA is the best in F1-OUT-SENTENCE (with FLAN-T5-LORA-HCN-NC closely behind) showcasing good document-level reasoning, presumably because it was trained in SQUAD and NQ.

6 Conclusions and Future Work

We propose an efficient approach for detecting *Claims*, *Claimers*, and *Claim Objects* using fine-tuned LLMs with LoRA. Our model is a FLAN-T5, fine-tuned in a new dataset of Health, Climate Change and Nutrition (HCN) news articles. We evaluate our model on HCN and compare it with baseline models on NEWSCLAIMS. Our approach outperforms the baselines in NEWSCLAIMS, is competitive with a QA model pretrained in SQUAD and NQ, and surpasses it in *Claimer* and *Claim Object* detection. In future work, we aim to explore more sophisticated LLMs, by also providing a comparative analysis amongst them and investigate methods for generating multiple claims from a single prompt. *Claim* detection serves as the initial step in claim verification. We plan to explore science-driven claim verification approaches using fine-tuned LLMs for extracting claims from scientific articles and utilize LLMs for polarity detection

between news article claims and scientific claims.

Limitations

In this section, we discuss some limitations of our work. As mentioned in Sections 4.1 and 4.2, the HCN dataset has more than one claimers and claim objects in each article. The same applies in NEWSCLAIMS, where we also have multiple claims and claim objects per article. One limitation of our fine-tuned model is that it is not capable of generating multiple outputs (e.g., multiple claims per article). We alleviate this, in the *claimer* and *claim object* detection by providing as many instances as are the claimers and claim objects. However, the problem remains for the claims. One solution, would be to create only one *instruction, answer* pair but separate the multiple answers with the token "|" or a similar token. This solution requires however further experimentation and is left for future work. To remedy this and be able to evaluate our model in NEWSCLAIMS, we proposed the scoring mechanism described in Section 4.2.

Nevertheless, the limitations of this approach are three-fold. Firstly, the aforementioned scoring mechanism favors choices with a lot of tokens, secondly if multiple choices have a lot of overlapping tokens, the scores will be smoothed out and thirdly, we must know beforehand the possible answers. As a result, we can not employ the scoring mechanism for the *claimer* and *claim object* tasks. The claimers are entities (e.g., persons, organizations), so a Named Entity recognition model is applicable, however the performance of the fine-tuned model depends on the performance of the Named Entity recognizer. Additionally, the claim objects are small snippets of text and it would require to generate n-grams for possible answers, with a lot of n-grams having overlapping in tokens.

Ethics Statement

The intended use of HCN is to evaluate methodological work regarding *claim*, *claim object* and *claimer object* detection in the domains of Health, Nutrition and most importantly in Climate Change, where to our knowledge no other dataset like this exists. HCN is not intended to directly make conclusions regarding the journalism quality nor quantify disagreement regarding the domains in the dataset. It is also worth mentioning that since the LLM was instruction fine-tuned with gold annotated claims, claimers and claim objects from the text and explicitly instructed to select snippets from the text (as it is shown in Figures 3,5,4, section A.3 in the Appendix, i.e. converting the task to an extractive QA task), it will always select a snippet from the news article, mitigating the hallucinations that accompany these generative models.

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback. This work was supported by research grants from the EEA and Norway Grants under the project UNBIASED (GR-INNOVATION-0009)[12], European Union's Horizon 2020 Research and Innovation Programme under grant agreement 952026[13] and European Union's Horizon Europe Research and Innovation Programme under grant agreement 101058573[14]. Finally, this work will also be part of the SCINOBO toolkit (Gialitsis et al., 2022; Kotitsas et al., 2023; Stavropoulos et al., 2023), developed by the Institute for Language and Speech Processing.

References

S Agrestia, AS Hashemianb, and MJ Carmanc. 2022. Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection. *Working Notes of CLEF*.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

[12] https://unbiasedproject.eu/

[13] https://cordis.europa.eu/project/id/952026

[14] https://cordis.europa.eu/project/id/101058573

Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Yang Zonghan, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:1–16.

Ahmet Bahadir Eyuboglu, Mustafa Bora Arslan, Ekrem Sonmezer, and Mucahid Kutlu. 2022. Tobb etu at checkthat! 2022: detecting attention-worthy and harmful tweets and check-worthy claims. *Working Notes of CLEF*.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

Nikolaos Gialitsis, Sotiris Kotitsas, and Haris Papageorgiou. 2022. Scinobo: A hierarchical multi-label classifier of scientific publications. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 800–809, New York, NY, USA. Association for Computing Machinery.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1835–1838.

Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10:1945–1948.

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

Zehra Melce Hüsünbeyi, Oliver Deck, and Tatjana Scheffler. 2022. Rub-dfl at checkthat!-2022: Transformer models and linguistic features for identifying relevant claims. In *Conference and Labs of the Evaluation Forum*.

Ria Jha, Ena Motwani, Nivedita Singhal, and Rishabh Kaushal. 2023. Towards automated check-worthy sentence detection using gated recurrent unit. *Neural Computing and Applications*, 35:1–21.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2020. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection.

Sotiris Kotitsas, Dimitris Pappas, Natalia Manola, and Haris Papageorgiou. 2023. Scinobo: a novel system classifying scholarly communication in a dynamically constructed hierarchical field-of-science taxonomy. *Frontiers in Research Metrics and Analytics*, 8.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering

research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. COVID-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–144, Dublin, Ireland. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. Scitab: A challenging benchmark for compositional reasoning and claim verification on scientific tables.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, and Alex Nikolov. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Ren Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard H. Hovy, Kevin Small, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with attribute knowledge. In *Conference on Empirical Methods in Natural Language Processing*.

Justin Reppert, Ben Rachbach, Charlie George, Luke Stebbing, Jungwon Byun, Maggie Appleton, and Andreas Stuhlmüller. 2023. Iterated decomposition: Improving science q&a by supervising reasoning processes.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2019. Pseudolikelihood reranking with masked language models. *CoRR*, abs/1910.14659.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering knowledge discovery from scientific literature: A novel approach to research artifact analysis. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 37–53, Singapore. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

A Appendix

A.1 HCN statistics

Table 5, presents statistics regarding the HCN claimers breakdown. We report on the number

of reported claimers, the number of Journalists (i.e. no reported claimer in the news article) and we also provide the breakdown in inside and outside of the major claim sentence. Furthermore, we observe that most of the times, the Journalist included the relevant sources outside of the major claim sentence. In cases where the Journalist did not, there was usually a plain artifact mentioned in the major claim sentence, such as “a report” or “a study”. In some cases with only generic references to the claimer, words such as “researchers” or “scientists” or “the study”, are hyperlinked to the source article itself. So, the Journalist has indicated a valid source of reference, outside of the text.

Claimer	Value
Reported Claimers	1369
Journalists	103
Claimers Inside Major Claim	175
Claimers Outside Major Claim	1194

Table 5: Number of Reported Claimers along with their breakdown in inside and outside of the major claim and number of journalists (i.e. when no reported claimer exists).

Figure 2, presents the verbs identified inside the major claims. Verbs such as "say", "suggest", "find" are observed in high numbers in the claim sentences, even though the majority of the claimers is identified outside of the major claim. Again, that happens because artifacts ("a report said", "a study showed") or plain nouns ("researchers said", "scientists found") accompany the verbs, and either a claimer is identified somewhere else in the text, or the Journalist does not include relevant sources, hence the Journalist is considered to be the claimer.

Figure 2: Wordcloud demonstrating the frequency of the verbs that are present in the major claim sentences.

In Table 6, we follow a similar approach with Table 5 and present the breakdown of the claim objects in inside and outside of the major claim.

Claim Object	Value
Claim Objects inside Major Claim	1101
Claim Objects outside Major Claim	112

Table 6: Number of Claim Objects inside and outside of the major claim sentence.

A.2 Instructions for fine-tuning

The instructions to perform fine-tuning and inference using LoRA are presented in Table 7.

A.3 Implementation Details

To train our foundation model FLAN-T5 in HCN, we fine-tune it with LoRA, using the HUGGINGFACE PEFT (Ding et al., 2023) library. Furthermore, for logging, visualizations and hyperparameter tuning we use the Weights & Biases (WANDB) library[15]. The hyperparameter tuning is done through a sweep in WANDB, using Bayesian optimization. The parameters we tune are r, LORA ALPHA, LR, EPOCHS, DROPOUT, where r is the number of ranks of the decomposition matrices, LORA ALPHA are the number of dimensions and LR is the learning rate. The possible values are the following: $r = \{4, 8, 16, 32\}$, LORA ALPHA = $\{8, 16, 32, 64\}$, DROPOUT = $\{0.1, 0.2, 0.25, 0.3, 0.35, 0.4\}$, EPOCHS = $\{5, 10, 15\}$ and LR is taking values from a uniform distribution from a range of $\{0.000005 - 0.0005\}$. We fine-tuned on a Quadro RTX 5000 of 16GB VRAM, with ADAM as the optimizer, employed a LR scheduler with warmup steps of 100. We optimized for the best token-wise F1 score for the claim detection task on the development set, since the claimers and claim objects are dependent on the major claim of a news article. We also employed early stopping of patience three. The best set of parameters on the development set were: $r = 4$, LORA ALPHA = 8, DROPOUT = 0.4, LR = 0.0002, EPOCHS = 15. We also used a maximum length of tokens for the FLAN-T5 of 1024. For fine-tuning in the NEWSCLAIMS dataset the same set of parameters were used, since we only had 18 articles to fine-tune our models and not enough for tuning the parameters again.

A.4 Qualitative Examples

To better understand, the inference procedure and the generative capabilities of a fine-tuned FLAN-T5 model, we will present some qualitative examples.

[15] https://wandb.ai/site/research

Sub-Task	Instruction
Claim detection	'What is the main claim of the input? Select a snippet of text from the input.'
Claimer detection	'Who is the claimer of the claim? Select a snippet of text from the input. If there is not a claimer in the input, write "No claimer found".'
Claim Object detection	'What is the topic of the claim? Select a snippet of text from the input.'

Table 7: Instructions per sub-task. The sub-tasks correspond to the claim, claimer and claim object detection.

True Label	Generated Output
'Roman Grüter', 'colleagues at Zurich University of Applied Sciences'	'roman grüter and colleagues at zurich university of applied sciences, switzerland'
'Scripps Institution of Oceanography at UC San Diego', 'University of Hawaii'	'scripps institution of oceanography at uc san diego, university of hawaii'
'Researchers from UCLA', 'the university of Washington'	'ucla, the university of washington'

Table 8: *Claimer* examples from the HCN dataset, where the news article has multiple *claimers*. In the column *True Label*, the multiple *claimers* are seperated with comma. In the column *Generated output*, the whole answer from FLAN-T5-LORA-HCN is visible.

True Label	Generated Output
Emissions	Carbon emissions reduction
Tree species	Tree diversity
Omicron	Omicron Variant
Cholesterol Reduction	Cholesterol
Limbs	Human Limbs
Researchers from UCLA	UCLA
The U.S agriculture secretary	The U.S. department of agriculture

Table 9: Instances where the true label with the generated output are contextually similar and correct. The last two rows are *claimers* and the rest of the rows are *claim objects*.

Recall from Section 4.1, that in the HCN dataset, even if we have one *major claim* per news article, we may have multiple claimers and claim objects. Our intuition is that by fine-tuning FLAN-T5 with multiple instances per claimer (claim object) for the same claim, the model will be able to generate a compositional answer containing the multiple claimers (claim objects) of the claim. Examples of this behaviour are presented in Table 8 for the claimer task. From the Table is visible that the model is capable of generating an answer that encapsulates both of the *claimers* present in each example in the column *True Label*. The negative side effect, is that when evaluating we treat the multiple *claimers* in the column *True Label* as two separate instances. Since we are calculating token-wise F1 (F1*), none of these instances will have 100% of F1*.

The aforementioned effect, also corresponds with the fact that the model might extract answers that are synonyms or contextually similar with the true labels, in the sub-tasks of *claimer* and *claim object* detection. These instances might count as wrong predictions of the model. Examples are pre-

sented in Table 9. From the presented Table, it is evident that all of these generated outputs are correct, however F1* will punish these predictions, deteriorating the reported performance. A solution would be to perform a round of human evaluation, assessing these predictions as in previous work (Tsatsaronis et al., 2015).

A.5 Input template examples

In Figures 3, 4 and 5, we can observe the template inputs along with an example. Notice that for the *claimer* and *claim object* sub-tasks we also provide the claim of the news article.

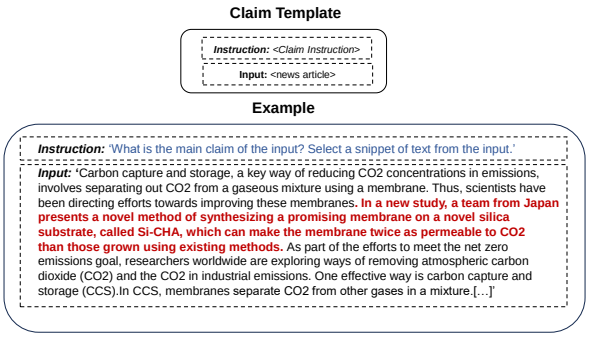

Figure 3: Claim Template that is used as input to the fine-tuned FLAN-T5 model. An example of a news article from the HCN dataset is also presented.

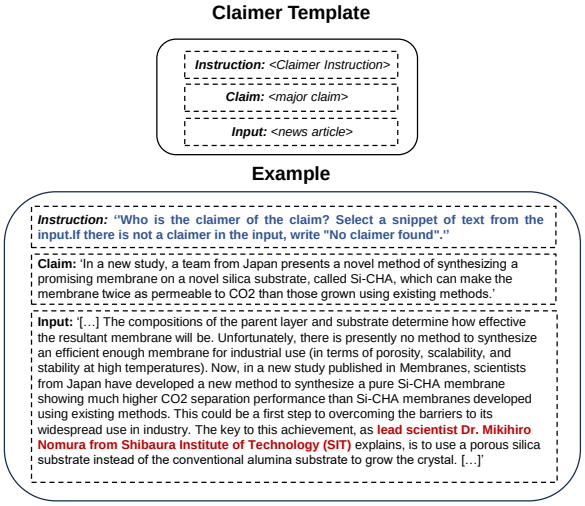

Figure 4: Claimer Template that is used as input to the fine-tuned FLAN-T5 model. An example of a news article from the HCN dataset is also presented.

A.6 HCN annotation examples

In this section of the Appendix, we present some examples from the HCN dataset along with their annotations (*claim*, *claimer*, *claim object*). Refer to

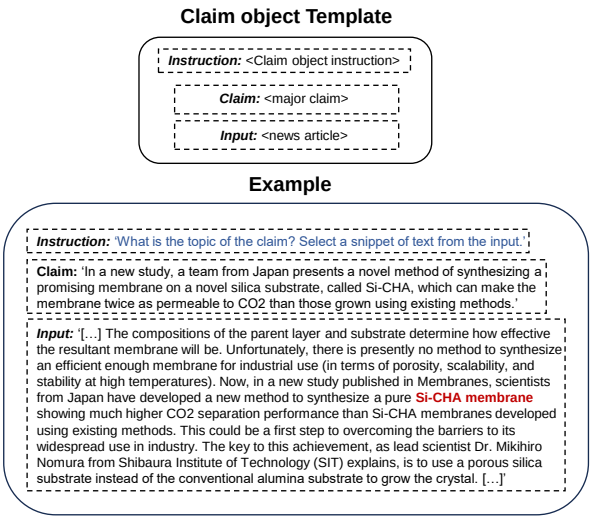

Figure 5: Claim Object Template that is used as input to the fine-tuned FLAN-T5 model. An example of a news article from the HCN dataset is also presented.

Figure 6. As it is evident from the Figure, the *claim object* in our HCN dataset is not always present in the *major claim* of the news article, inherently making the sub-task of *claim object* detection more challenging for the HCN dataset.

Urban greening is unlikely to provide a single fix for tackling extreme weather events brought on by climate change, scientists have suggested. A team led by researchers from Cardiff University has shown that the majority of cities around the world will not be able to reduce instances of heatwaves and flooding at the same time through the introduction of strategies such as green roofs, living walls, vegetated urban spaces and parks. Publishing their findings today in the journal Nature Communications, the team show that the cooling or flood-reducing potential of green urban spaces depends strongly on the prevailing climate of the city in question, with flood protection likely to be more successful in arid environments, whilst a cooling effect more likely in more humid climates. […]

Major Claim: A team led by researchers from Cardiff University has shown that the majority of cities […]
Claimer: Researchers from Cardiff University
Claim Object: Urban Greening

A new analysis predicts that, as climate change progresses, the most suitable regions for growing coffee arabica, cashews, and avocados will decline in some of the main countries that produce these crops. Roman Gruter and colleagues at Zurich University of Applied Sciences, Switzerland, present these findings in the open-access journal PLOS ONE on January 26, 2022. […]

Major Claim: A new analysis predicts that, as climate change progresses […]
Claimer: Roman Gruter, Colleagues at Zurich University of Applied Sciences, Switzerland
Claim Object: Crops

A new study from North Carolina State University shows that methane, a potent greenhouse gas, is largely generated in the soils below standing dead trees in so-called ghost forests, or coastal forests that are being killed off by rising sea levels. This escaping methane gas, known colloquially as ghost forest tree "farts" is actually generated by different tiny microorganisms. Researchers wanted to know if different communities of microbes are making methane gas inside the soils or in the dead trees, which are also known as snags. […]

Major Claim: A new study from North Carolina State University shows that methane, a potent greenhouse gas, is largely generated in the soils […]
Claimer: North Carolina State University
Claim Object: Methane

By 2080, around 70% of the world's oceans could be suffocating from a lack of oxygen as a result of climate change, potentially impacting marine ecosystems worldwide, according to a new study. The new models find mid-ocean depths that support many fisheries worldwide are […] "This zone is actually very important to us because a lot of commercial fish live in this zone" says Yuntao Zhou, an oceanographer at Shanghai Jiao Tong University and lead study author. […]

Major Claim: By 2080, around 70% of the world's oceans could be suffocating from a lack of oxygen as a result of climate […]
Claimer: Yuntao Zhou, an oceanographer at Shanghai Jiao Tong University and lead study author.
Claim Object: Climate Change

Women may rest a bit easier thanks to results from a study showing that coronavirus vaccines have almost no impact on a woman's menstrual cycle. The issue is significant, as regular menstruation is a sign of health and fertility, and fears of disturbances could make people less likely to get a vaccine as COVID-19 cases continue to surge. […] Alison Edelman, MD, a professor of obstetrics and gynecology at Oregon Health & Science University, led a group studying data […]

Major Claim: Women may rest a bit easier thanks to results from a study showing that coronavirus vaccines […]
Claimer: Alison Edelman, MD, a professor of obstetrics and gynecology at Oregon Health & Science University
Claim Object: Menstrual cycle

Belly fat is usually unwelcome, but new research suggests it may actually be good for something: relief from foot pain. A small pilot study suggests that an injection of a patient's own fat cells can help ease the often-excruciating heel pain brought on by a condition known as plantar fasciitis. "We take a small amount of fat from an area of excess like the belly, inner thigh or love handles and then inject the fat into the bottom of the foot near where the fascia inserts into the heel bone" explained study co-author Dr. Jeffrey Gusenoff […]

Major Claim: A small pilot study suggests that an injection of a patient's own fat cells […]
Claimer: Study co-author Dr. Jeffrey Gusenoff
Claim Object: plantar fasciitis

Figure 6: Annotation examples from the HCN dataset. With green color is the *claim object*, with red color is the *major claim* and with blue color is the *claimer*.