QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims

Venktesh V
v.Viswanathan-1@tudelft.nl
Delft University of Technology
Delft, Netherlands

Abhijit Anand
aanand@L3S.de
L3S Research Center
Hannover, Germany

Avishek Anand
avishek.anand@tudelft.nl
Delft University of Technology
Delft, Netherlands

Vinay Setty*
vsetty@acm.org
University of Stavanger
Stavanger, Norway

ABSTRACT

With the growth of misinformation on the web, automated fact checking has garnered immense interest for detecting growing misinformation and disinformation. Current systems have made significant advancements in handling synthetic claims sourced from Wikipedia, and noteworthy progress has been achieved in addressing real-world claims that are verified by fact-checking organizations as well. We compile and release **QuanTemp**, a diverse, multi-domain dataset focused exclusively on numerical claims, encompassing comparative, statistical, interval, and temporal aspects, with detailed metadata and an accompanying evidence collection. This addresses the challenge of verifying real-world numerical claims, which are complex and often lack precise information, a gap not filled by existing works that mainly focus on synthetic claims. We evaluate and quantify these gaps in existing solutions for the task of verifying numerical claims. We also evaluate claim decomposition based methods, numerical understanding based natural language inference (NLI) models and our best baselines achieves a macro-F1 of 58.32. This demonstrates that QuanTemp serves as a challenging evaluation set for numerical claim verification.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Fact-checking, Numerical Claims, Claim Decomposition

ACM Reference Format:
Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657874

*Also with Factiverse AI.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657874

Example: Claim from QuanTemp

Claim: `Under GOP plan, U.S. families making $86k see avg tax increase of $794.`

[Evidence]: If enacted, the Republican tax reform proposal would saddle only 8 million households that earn up to $86,100 with an average tax increase of $794 .... Only a small percentage (6.5 percent) of the nearly 122 million households in the bottom three quintiles will actually face a tax increase.

[Verdict]: False

Figure 1: Example claim from QuanTemp

1 INTRODUCTION

Online misinformation, particularly during elections, poses a significant threat to democracy by inciting socio-political and economic turmoil [49]. Fact-checking websites like `Politifact.com`, `Snopes.com`, `FullFact.org`, and others play an indispensable role in curbing misinformation. However, the scalability of manual fact-checking is constrained by limited resources. This limitation has spurred remarkable advancements in neural models for automated fact-checking in recent years [15], driven by the proliferation of open datasets [7, 11, 30, 39, 45]. Crucially, within the area of fact-checking, the verification of claims involving numerical quantities and temporal expressions is of utmost importance. This is essential for countering the 'numeric-truth-effect'[35], where the presence of numbers can lend a false aura of credibility to a statement. Numerical claims are a significant component of political discourse. For instance, our analysis of the ClaimBuster dataset [16] reveals that a substantial 36% of all check-worthy claims in U.S. presidential debates involve numerical quantities. Most current datasets inadequately address the verification of numerical claims, as our overview in Table 1 illustrates. A notable example is the Feverous dataset, where only a small fraction (approximately 10%) of claims necessitate numerical reasoning, and these have proven especially challenging for annotators to verify [2]. Our experiments further reinforce this difficulty. We observed that models trained on a mix

of numerical and non-numerical claims underperform compared to those specifically fine-tuned on numerical claims.

Numerical claims verification poses a unique challenge, where a fact-checking system must critically analyze and reason about the numerical data presented in both the claim and its evidence. For example, in verifying the claim shown in Figure 1 as 'False', the NLI model needs to identify that the evidence only mentions 8 million households with incomes up to $86k facing tax increases, contradicting the claim of tax increases for all families earning $86k. The existing datasets can be categorized as synthetically generated from Wikipedia and knowledge bases or real-world claims collected from fact-checking websites (Table 1). While, works like ClaimDecomp [10], MultiFC [7] and AveriTec [39], collect real-world claims, they do not particularly focus on numerical claims. The only previous work proposing a dataset for fact-checking statistical claims, by [43, 47], uses a rule-based system to create synthetic claims from 16 simple statistical characteristics in the Freebase knowledge base about geographical regions. There has not been a dedicated large-scale real-world open-domain diverse dataset for verifying numerical claims.

In this work, we collect and release a dataset of 15, 514 real-world claims with **numeric quantities and temporal expressions** from various fact-checking domains, complete with detailed metadata and an evidence corpus sourced from the web. *Numeric claims are defined as statements needing verification of any explicit or implicit quantitative or temporal content.* The evidence collection method is crucial in fact-checking datasets. While datasets like MultiFC and DeClarE use claims as queries in search engines like Google and Bing for evidence, methods like ClaimDecomp depend on fact-checkers' justifications. However, this could cause 'gold' evidence leakage from fact-checking sites into the training data. To avoid this, we omit results from fact-checking websites in our evidence corpus.

Moreover, using claims as queries in search engines may miss crucial but non-explicit evidence for claim verification. To overcome this, recent works have proposed generating decomposed questions to retrieve better evidence [3, 10, 14, 26, 31]. In our approach, we aggregate evidence using both original claims and questions from methods like ClaimDecomp and ProgramFC. This dual strategy yields a more diverse and unbiased evidence set of **423,320 snippets**, enhancing which could be used for evaluating both the retrieval and NLI steps of fact-checking systems.

Finally, we also propose a fact-checking pipeline as a baseline that integrates claim decomposition techniques for evidence retrieval, along with a range of Natural Language Inference (NLI) models, encompassing pre-trained, fine-tuned, and generative approaches, to evaluate their efficacy on our dataset. Additionally, we conduct an error analysis, classifying the numerical claims into distinct categories to better understand the challenges they present.

1.1 Research Questions

In addition to collecting and releasing the dataset, we answer the following research questions by proposing a simple baseline system for fact-checking.

RQ1: How hard is the task of fact-checking numerical claims?

RQ2: To what extent does claim decomposition improve the verification of numerical claims?

RQ3: How effectively do models pre-trained for numeric understanding perform when fine-tuned to fact-check numerical claims?

RQ4: How does the size of large language models impact their performance in zero-shot, few-shot, and fine-tuned scenarios for numerical claims?

1.2 Contributions

(1) We collect and release a large, diverse multi-domain dataset of real-world **15,514 numerical claims**, the first of its kind, along with an associated evidence corpus consisting of **423,320** snippets.
(2) We evaluate established fact-checking pipelines and claim decomposition methods, examining their **effectiveness in handling numerical claims**. Additionally, we propose improved baselines for the natural language inference (NLI) step.
(3) Our findings reveal that NLI models pre-trained for **numerical understanding outperform generic models** in fact-checking numerical claims by up to 11.78%. We also show that smaller models fine-tuned on numerical claims outperform larger models like GPT-3.5-Turbo under zero-shot and few-shot scenarios.
(4) We also **assess the quality of questions** decomposed by ClaimDecomp and ProgramFC for numerical claims, using both automated metrics and manual evaluation.

We make our *dataset and code* available here https://github.com/factiverse/QuanTemp.

2 RELATED WORK

The process of automated fact-checking is typically structured as a pipeline encompassing three key stages: claim detection, evidence retrieval, and verdict prediction, the latter often involving stance detection or natural language inference (NLI) tasks [8, 15, 56]. In this work, we introduce a dataset of numerical claims that could be used to evaluate the evidence retrieval and NLI stages of this pipeline. This section will explore relevant datasets and methodologies in this domain.

Most current fact-checking datasets focus on textual claims verification using structured or unstructured data [44, 56]. However, real-world data, like political debates, frequently involve claims requiring numerical understanding for evidence retrieval and verification. It has also been acknowledged by annotators of datasets such as FEVEROUS that numerical claims are hard to verify since they require reasoning and yet only 10% of their dataset are numerical in nature [2].

A significant portion of the existing datasets collect claims authored by crowd-workers from passages in Wikipedia [2, 19, 29, 38, 40, 45]. Additionally, there are synthetic datasets that require tabular data to verify the claims [2, 11, 24], but these claims and tables may not contain numerical quantities. Recent efforts by [20] aim to create more realistic claims from Wikipedia by identifying cited statements, but these do not reflect the typical distribution of

Table 1: Comparison of QuanTemp with other fact checking datasets. *Some datasets may have some numerical claims in them, but it is not their main focus. †By "Unleaked Evidence", here we refer to gold evidence being leaked from fact-checking websites. In the table WP refers to Wikipedia, WT refers to WikiTables, FCS referes to fact-check sites and FB referes to FreeBase

Dataset	# of claims	Claims Source	Retrieved Evidence	Numerical Focus*	†Unleaked Evidence†	Evidence Corpus Size
Synthetic Claims						
StatProps [43]	4,225	KB (FB)	KB (FB)	✓	N/A	N/A
Fever [45]	185,445	WP	WP	✗	N/A	5,416,568
Hover [19]	26,171	WP	WP	✗	N/A	5,486,211
TabFact [11]	92,283	WP	WT	✗	N/A	16,573
Feverous [2]	87,026	WP	WT,WP	✗	N/A	7,221,406
SciTab '[24]	1,225	arXiv (CS)	arXiv	✗	N/A	1,301
Fact-checker Claims						
LIAR [52]	12,836	Politifact	✗	✗	N/A	N/A
ClaimDecomp [10]	1,250	Politifact	✗	✗	✗	1,250
DeClarE [30]	13,525	FCS (4)	Web	✗	✓	87,643
MultiFC [7]	36,534	FCS (26)	Web	✗	✗	349,180
QABriefs [14]	8,784	FCS (50)	Web	✗	✗	21,168
AVeriTeC [39]	4,568	FCS (50)	Web	✗	✗	137,040
QuanTemp (OURS)	15,514	FCS (45)	Web	✓	✓	423,320

claims verified by fact-checkers and the false claims they contain are still synthetic.

More efforts have been made to collect real-world claims in domains like politics [1, 10, 27, 52], science [48, 50, 54], health [22] and climate [13] and other natural claims occurring in social media posts [12, 25]. Multi-domain claim collections like MultiFC [7] have also emerged, offering rich meta-data for real-time fact-checking. However, none of these datasets focus on numerical claims.

Among those that focus on numerical claims, [43, 47], the authors propose a simple distant supervision approach using freebase to verify simple statistical claims. These claims are not only synthetic, but they can be answered with simple KB facts such as Freebase. Similarly, [9] explore the extraction of formulae for checking numerical consistency in financial statements by also relying on Wikidata. Further, [18] explore the identification of quantitative statements for fact-checking trend-based claims. None of these datasets are representative of real-world claims.

In this work, we collect and release a multi-domain dataset which is primarily composed of numerical claims and temporal expressions with fine-grained meta-data from fact-checkers and an evidence collection. To the best of our knowledge, this is the first natural numerical claims dataset. Another set of related work is the area of temporal information retrieval that deals exclusively with time [21, 41]. However, these works do not focus on temporal reasoning and instead deal with indexing and processing temporal queries [4–6, 17].

Early fact checking systems simply used the claim as the query to search engine [7, 15, 30] or employ question answering systems [36]. These approaches may not work well if the claims are not already fact-checked. In this regard, recent works have introduced claim decomposition into questions [10, 14, 28, 39]. In this paper, we evaluate the effectiveness of ClaimDecomp [10] and ProgramFC [28] methods for numerical claims.

We follow the established fact-checking pipeline using evidence and claims as input to NLI models to predict if the claims are supported, refuted or conflicted by the evidence [15]. We use BM25 [34] for evidence retrieval followed by re-ranking and explore various families of NLI models.

3 DATASET CONSTRUCTION

In this section, we describe an overview of the dataset collection process as shown in Figure 2.

3.1 Collecting Real-world Claims

We first collect real-world claims curated by professional fact-checkers at fact-checking organizations via Google Fact Check Tool APIs[1]. The full collection of fact-checks consists of 278,636 fact-checks[2]. After filtering non-English fact-checks, it amounts to 45 organizations worldwide with 139,926 fact-checks. Next, we identify quantitative segments (Section 3.3) from the claims and only retain claims that satisfy these criteria, which amounts to 15,514 claims. Finally, we collect evidence for the claims (Section 3.5).

One of the challenges of collecting claims from diverse sources is the labelling conventions. To simplify, we standardize the labels to one of *True, False or Conflicting* by mapping them similar to [39]. We also ignore those claims with unclear or no labels.

3.2 Dataset Statistics

After deduplication, our dataset has **15,514** claims. The dataset is unbalanced, favoring refuted or false claims with a distribution

[1]https://toolbox.google.com/factcheck/apis available under the CC-BY-4.0 license.
[2]As of 15th Nov 2023.

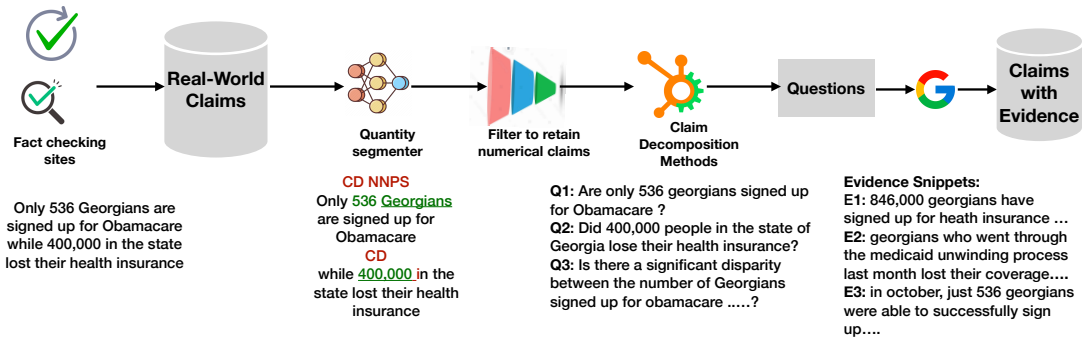

Figure 2: QuanTemp Construction Pipeline

Table 2: Top fact-checking domains

Claim Source	#Occurences
Politifact	3,840
Snopes	1,648
AfP	412
Africacheck	410
Fullfact	349
Factly	330
Boomlive_in	318
Logically	276
Reuters	235
Lead Stories	223

Table 3: Top claim source countries.

Country	#Occurences
USA	6,215
India	1,356
UK	596
France	503
South Africa	410
Germany	124
Philippines	103
Australia	65
Ukraine	35
Nigeria	17

Table 4: Top evidence domains.

Category	#Occurences
en.wikipedia.org	28,124
nytimes.com	8,430
ncbi.nlm.nih.gov	8,417
quora.com	4,967
cdc.gov	3,987
statista.com	3,106
youtube.com	2,889
who.int	2,557
cnbc.com	2,448
investopedia.com	1977

Table 5: QuanTemp dataset distribution

Split	True	False	Conflicting	Total
Train	1,824	5,770	2,341	9,935
Dev	617	1,795	672	3,084
Test	474	1,423	598	2,495

of 'True', 'False', and 'Conflicting' claims at 18.79%, 57.93%, and 23.27%, respectively, reflecting the tendency of fact-checkers to focus on false information [39]. A detailed distribution of the dataset is shown in Table 5.

To illustrate the comprehensive nature of our dataset, we analyze on the origins of the claims it encompasses. Table 2 presents a summary of the ten most frequently encountered fact-checking websites within our dataset. It is noteworthy that Politifact constitutes a substantial fraction of the claims. Additionally, our analysis reveals a significant representation of claims sourced from a variety of fact-checking platforms, extending beyond the realm of political fact-checks. Furthermore, as illustrated in Table 3, our dataset encompasses claims originating from a wide range of geographical locales, including, North America, Europe, Asia, and Africa, underscoring the geographical diversity of our corpus.

3.3 Identifying Quantitative Segments

We identify quantitative segments in the claim sentence for extracting numerical claims, as defined in [32], which include numbers,

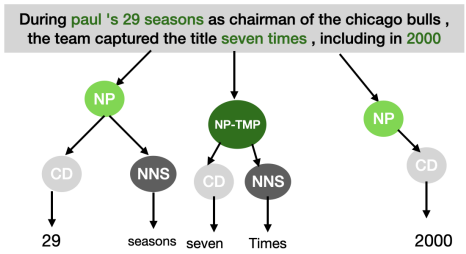

Figure 3: Example of identification of quantitative segments from the claim. NNS is noun plural form, NP-TMP is temporal noun phrase

units, and optionally approximators (e.g.,"roughly") or trend indicators (e.g., "increases").

Specifically, we first obtain the claim's constituency parse, identifying nodes with the cardinal number POS tag "CD". To avoid false positives (for example: "The one and only"), we then parse these nodes' ancestors and extract noun phrases from their least common ancestors. Using these noun phrases as root nodes, we perform a prefix traversal of their subtrees. Figure 3 shows an example of the extracted quantitative segments. We then refine the claim set by filtering for those with at least one quantitative segment.

This approach is limited, as it may include claims with non-quantitative terms like "Covid-19". To remedy this, we require more

than one quantitative segment, excluding any nouns like "Covid-19" mentions, to qualify as a numerical claim. Claims not meeting this criterion are excluded. Our self-assessment of 1000 sample claims from the dataset indicates a 95% accuracy of our heuristic.

3.4 Claim Decomposition

We posit that the precise nature of the task of verifying numerical information requires the retrieval of relevant evidence containing the required quantitative information. Claim decomposition [10, 14, 26, 28] is effective in extracting important evidence containing background or implied information to verify the claim. Hence, we evaluate several claim decomposition approaches as described in Section 4.1 and discuss their impact on downstream veracity prediction in Section 5.4. We also release the relevant questions resulting from the decomposition of the claim for train, validation, and test sets to facilitate further research on their significance for retrieving relevant evidence.

3.5 Collecting Evidence

We collect evidence for claim verification by retrieving search results from public engines, following prior works [7, 39]. Solely using the claim as a query, as in prior approaches [7, 8, 30], yields a restricted set of documents, often biased towards claims that have already been verified. To enhance the diversity of the evidence, we include both the original claim and questions generated from the decomposition methods, described in Section 4.1. We get inspiration from recent advances in claim decomposition methods for fact-checking that have shown promising results [14, 28, 39]. To ensure the evidence collection is free from duplicates, we merge the search results for the original claim and all the generated questions using a top-k document pooling method commonly used in information retrieval (IR) [37, 42].

We first submit original claims to Google through scaleserp.com API, collecting the top 10 results per claim. We strictly filter out any results from over 150 fact-checking domains to prevent any leakage from their justification, avoiding models to learn shortcuts in verification. We improve evidence diversity by employing Large Language Model (LLM)-based claim decomposition methods such as Program-FC [28] and ClaimDecomp [10], generating a range of questions for query use. For each question, we compile the top 10 search results.

After filtering out duplicates and irrelevant documents, we amass an extensive evidence collection of 423,320 snippets. We have evidence from a diverse set of domains including Wikipedia, government websites, etc. An overview of top evidence domains is shown in Table 4. We ensure our collection does not have any snippets from manual or automated fact-checkers and related websites or social media handles. Also, government websites are one of the frequently occurring domains in our evidence collection, as our claims comprise diverse political and international events. Not surprisingly, many evidence snippets are from `statisa.com` and `investopedia.com` since our dataset is focused on numerical claims. On the other hand, we have 2,889 snippets from `youtube.com` which is surprising. After closer inspection, these snippets contain transcripts from news videos and textual descriptions which that relevant to the claim.

3.6 Qualitative Analysis of the Dataset

To ensure that the dataset we collect is of good quality, we perform automated and manual evaluations. Specifically, in manual evaluation, we evaluate the usefulness and comprehensiveness of the generated questions by various claim decomposition methods and the usefulness of the evidence they yield towards verifying the claim. We detail the guidelines used in the annotation process in this section.

We rate the questions generated on 2 aspects: completeness and usefulness (only based on the claim given and top search results). The two annotators are trained computer scientists who are closely associated with the task of automated fact checking and are familiar with the domain. The following guidelines were provided to the annotators.

Completeness: A list of questions is said to be complete if the questions cover all aspects of the claim. The ratings were carried out according to the Likert scale of 1-5.

Usefulness: Rating the usefulness of questions determines if the questions would help verify the claim. The annotators were asked to rate the usefulness of questions resulting from the decomposition of claims. They were instructed to consider implicit aspects of the claim when rating this, and were asked to rate according to the Likert scale 1-5.

Some questions might be relevant, but may just retrieve background knowledge and may not be relevant to the core aspect being fact checked. To gauge this, the annotators were also asked to look at evidence when rating the usefulness of questions.

While evaluating the usefulness, the annotators were asked not to make assumptions about verification method. They were instructed to check if the questions covered all aspects of the claim and could retrieve relevant evidence. They were also asked to check the coverage of the implied meaning of the claim, rather than just a surface level analysis of the claim. For instance, the implied aspect in the claim "The ICE spends 4 times the amount to detain a person for a year than on a student in public school. " is that ICE detains people for more than a year, which is not the case in reality. To verify this core aspect of the claim, the questions must be useful to retrieve related evidence that cites the time usually people are detained by the ICE.

Evidence Usefulness: The annotators were requested to rate the individual piece of evidences for each question. The usefulness depends on the information contained in the evidence. The information should be sufficient to support the whole or parts of the claim and should be rated on the Likert scale of 1-5 based on the degree of information. They were asked to rate the usefulness of all evidence by aggregating the usefulness of evidence tied to individual questions. The annotators were asked to rate individual evidence based on relatedness to the claim and their utility in fact checking.

4 EXPERIMENTAL SETUP

To evaluate the QuanTemp dataset, we introduce a baseline fact-checking pipeline. We fix the retriever model (BM25 + re-ranking) for all experiments. After extensive experiments, we choose to fine-tune the Roberta-Large-MNLI[3] model, pre-fine-tuned on the MNLI

[3]https://huggingface.co/roberta-large-mnli

corpus, for the NLI task. In Section 4.4 we further explore various NLI models' effectiveness on numerical claims.

4.1 Claim Decomposition

We evaluate the following claim decomposition approaches.

ClaimDecomp [10]: The authors of this work provide annotated yes/no sub-questions for the original claims. We use GPT-3.5-TURBO on training samples from the ClaimDecomp dataset to create yes/no sub-questions for our QuanTemp dataset through in-context learning, setting temperature to 0.3, frequency to 0.6 and presence penalties to 0.8.

Program-FC [28]: We implement the approach proposed in this work to decompose claims and generate programs to aid in verification. The programs are step by step instructions resulting from decomposition of original claim. We employ gpt-3.5-turbo for decomposition. We use same hyperparameters as in the original paper.

Original Claim: Here we do not employ any claim decomposition, but rather use the original claim to retrieve evidence for arriving at the final verdict using the NLI models.

Table 6: A broad overview of different categories of claims in QuanTemp

Category	Examples	#of claims
Statistical	We've got 7.2% unemployment (in Ohio), but when you include the folks who have stopped looking for work, it's actually over 10%.	7302 (47.07%)
Temporal	The 1974 comedy young frankenstein directly inspired the title for rock band aerosmiths song walk this way	4193 (27.03%)
Interval	In Austin, Texas, the average homeowner is paying about $1,300 to $1,400 just for recapture, meaning funds spent in non-Austin school districts	2357 (15.19%)
Comparison	A vaccine safety body has recorded 20 times more COVID jab adverse reactions than the government's Therapeutic Goods Administration.	1645 (10.60%)

4.2 Evidence retrieval and Veracity Prediction

Once we have decomposed claims, we use evidence retrieval and re-ranking. Then we employ a classifier fine-tuned on QuanTemp for the NLI task to verify the claim. The different settings in which we evaluate the approaches are:

Unified Evidence: Our experiments utilize the evidence snippets collection detailed in Section 3.5. For each question/claim, we retrieve the top-100 documents using BM25 and re-rank them with *paraphrase-MiniLM-L6-v2* from sentence-transformers library [33], selecting the top 3 snippets for the NLI task. We experiment with different values of $k \in [1, 3, 5, 7]$ for top-k snippets. We observe that k=3 at claim level works best for our setup. Hence, for the "Original Claim" baseline, we use the top 3 evidences using the claim, and for other methods, we use the top-1 evidence per question, ensuring three evidences per claim for fair comparison. We experiment with different encoder models in huggingface such as Deberta-base, Deberta-large, Roberta family of models and observe that the 6-layer MiniLM model [51] (*paraphrase-MiniLM-L6-v2*) provides the best evidence ranking for downstream fact-verification.

After retrieving evidence, we fine-tune a three-class classifier for veracity prediction. Training and validation sets are formed using the retrieved evidence (described above), and the classifier is fine-tuned by concatenating the claim, questions, and evidence with separators, targeting the claim veracity label.

Gold Evidence: Here, we directly employ the justification paragraphs collected from fact-checking sites as evidence to check the upper bound for performance.

Hyperparameters: All classifiers are fine-tuned till "EarlyStopping" with patience of 2 and batch size of 16. AdamW optimizer is employed with a learning rate of $2e-5$ and ϵ of $1e-8$ and linear schedule with warm up. We use transformers library [53] for our experiments.

4.3 Category Assignment

After curating the numerical claims, we categorize the numerical claims to one of these categories using a weak supervision approach. We identify four categories: temporal (time-related), statistical (quantity or statistic-based), interval (range-specific), and comparison (requiring quantity comparison) claims. The examples and distribution of these categories are shown in Table 6. We perform this categorization to aid in a fine-grained analysis of performance across different dimensions of numerical claims, as described in Section 5.6. The categorization would help understand the performance of existing models on different types of numerical claims and gauge the scope for improvement.

We first manually annotate 50 claims, then used a few samples as in-context examples for the gpt-3.5-turbo model to label hundreds more. After initial labeling, we fine-tuned Setfit [46], a classifier ideal for small sample sizes on the annotated samples. Then we employed the classifier for further annotation of the entire dataset to the defined categories. Two annotators manually reviewed 250 random claims, and 199 claims were found to be correctly categorized. These annotations helped refine the classifier and category assignment.

4.4 Veracity Prediction Model Ablations

Prompting based Generative NLI models: We assess the stance detection using large generative models like FLAN-T5-XL (3B params) and GPT-3.5-TURBO, providing them with training samples, ground truth labels, and retrieved evidence as in-context examples for claim verification. The models are also prompted to produce claim veracity and justification jointly to ensure faithfulness, with a temperature setting of 0.3 to reduce randomness in outputs. The few-shot prompt employed for veracity prediction through GPT-3.5-TURBO is shown in Table 12. We dynamically select few shot examples for every test example. The examples shown are for a single instance and are not indicative of the examples used for inference over all test

Table 7: Results of different models on QuanTemp (categorical and full) with Roberta-Large-MNLI as the NLI model. M-F1 : Macro-F1, W-F1 : Weighted-F1 and C-F1 refers to F1 score for Conflicting class.† indicates statistical significance at 0.01 level over baseline using t-test

Method	Statistical		Temporal		Interval		Comparison		Per-class F1			QuanTemp	
	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	T-F1	F-F1	C-F1	M-F1	W-F1
Unified Evidence Corpus													
Original Claim (baseline)	49.55	52.48	60.29	74.29	48.84	57.93	40.72	39.66	51.59	70.60	35.27	52.48	58.52
Program-FC	52.24	57.83	56.75	75.46	47.09	61.88	49.02	48.07	47.42	79.46	33.40	53.43	62.34
ClaimDecomp	53.34	58.79	61.46	78.02	56.02	66.97	53.59	53.44	51.82	79.82	39.72	**57.12** †	**64.89**†
Fine-tuned w/ Gold Evidence													
QuanTemp only	60.87	65.44	66.63	81.11	58.35	69.56	60.74	60.36	56.86	82.92	48.79	**62.85**	**69.79**
QuanTemp + non-num	56.76	61.98	64.04	80.35	56.56	67.13	52.03	50.59	59.87	83.13	33.78	58.66	66.73
Naive (Majority class)	22.46	34.25	28.35	62.95	25.86	49.19	16.51	16.32	0.00	72.64	0.00	24.21	41.42

examples. We also show the results for zero-shot veracity prediction using generative models.

4.4.1 Fine-tuned models. We fine-tune T5-small (60 M params), bart-large-mnli and Roberta-large (355 M params) to study the impact of scaling on verifying numerical claims. We also employ models pre-trained on number understanding tasks such as FinQA-Roberta-Large [57], NumT5-small [55]. We fine-tuned these models on our dataset for the NLI task to test the hypothesis if models trained to understand numbers better aid in verifying numerical claims. All models are fine-tuned with hyperparameters described in Section 4.2.

5 RESULTS

5.1 Automated Analysis of Quality of decomposition

Examining questions generated by ClaimDecomp and Program-FC, we prioritize their relevance to the original claim and diversity in covering different claim aspects. BERTScore [58] is employed to assess relevance, i.e., measuring how well the questions align with the claims. For diversity, which ensures non-redundancy and coverage of various claim facets, we utilize the sum of (1-BLEU) and Word Position Deviation [23]. The results are shown in Table 10. Our findings indicate that ClaimDecomp excels in generating questions that are not only more relevant but also more diverse compared to Program-FC, addressing multiple facets of the claim. We also perform manual analysis by sampling 20 claims sampled from test set along with decomposed questions and retrieved evidence for Program-FC and ClaimDecomp approaches.

5.2 Manual Evaluation Results

We perform a manual evaluation of questions from decomposed claims and retrieved evidence in QuanTemp as described in Section 3.6. We ask two computer scientists familiar with the field to annotate them on measures of completeness (if questions cover all aspects of the claim), question usefulness and evidence usefulness, where usefulness is measured by information they provide to verify the claim. The results are shown in Table 9. We also report the Cohen's kappa scores, indicating the level of agreement among the annotators in the table. We observe that questions generated by ClaimDecomp are better than Program-FC as measured by their usefulness and quality of evidence retrieved. We also observe that the decompositions yielded by both the approaches cover all aspects of the claim, as observed by the completeness score. We also release the questions from claim decomposition for train, validation, and test sets of QuanTemp.

5.3 Hardness of Numerical Claim Verification

To address **RQ1**, we experiment with various claim verification approaches on the QuanTemp dataset, considering both unified evidence and gold evidence. The performance of different approaches is presented in Table 7. *Fine-tune using Gold Evidence:* shows performance of NLI models fine-tuned on QuanTemp (numerical only) and QuanTemp+non-num (numerical and non-numerical claims) using gold evidence snippet. Both numerical claims and non-numerical claims are from the same fact-checkers.

It is evident that QuanTemp poses a considerable challenge for fact-checking numerical claims, with the best approach achieving a weighted-F1 of 64.89 for unified evidence and 69.79 for gold evidence. The difficulty is further underscored by the performance of the naive baseline, which simply predicts the majority class. A similar trend is observed at the categorical level. Except for the temporal category, where it outperforms other categories, the improvements from the baseline is relatively modest.

Additionally, training specifically on QuanTemp's numerical claim distribution improves performance by 7.14% in macro F1 compared to a mixed distribution of numerical and non-numerical claim set. These results underscore the complexity of verifying numerical claims.

5.4 Effect of Claim Decomposition on Claim Verification

The **RQ2** is answered by Table 7 which indicates that claim decomposition enhances claim verification, particularly for the 'Conflicting' category, where ClaimDecomp outperforms original claim-based verification significantly. In the 'unified evidence' setting,

Table 8: Ablation results employing different NLI models for ClaimDecomp on QuanTemp. The best results are in bold.† indicates statistical significance at 0.01 level over Roberta-large using t-test

Method	Statistical		Temporal		Interval		Comparison		Per-class F1			QuanTemp	
	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	T-F1	F-F1	C-F1	M-F1	W-F1
Unified Evidence Corpus													
bart-large-mnli	52.89	58.43	**62.01**	**78.07**	54.52	65.85	53.63	53.49	51.23	79.56	39.37	56.71	64.54
Roberta-large	53.56	58.24	59.31	75.67	51.64	62.38	43.91	42.39	50.58	77.23	35.50	54.43	62.16
T5-small	43.52	51.37	32.08	64.65	43.64	60.38	48.41	48.88	19.65	77.22	38.02	44.96	56.89
NumT5-small	50.36	56.58	41.35	68.59	47.96	61.35	49.14	48.90	36.56	78.45	35.76	50.26	60.26
FinQA-Roberta-Large	**56.97**	**61.36**	60.29	75.55	**56.53**	**66.52**	52.53	52.34	49.72	77.91	**47.33**	**58.32**†	**65.23**†
FlanT5 (zero-shot)	32.11	36.43	26.51	42.64	32.36	44.03	27.48	24.95	36.35	52.56	3.15	30.68	37.64
FlanT5 (few-shot)	37.31	41.24	32.70	46.83	37.61	47.52	35.20	34.47	33.90	54.73	20.92	36.52	42.67
gpt-3.5-turbo (zero-shot)	34.51	33.78	28.04	29.12	35.34	36.98	40.31	40.45	37.81	32.57	31.25	33.87	33.25
gpt-4 (few-shot)	37.04	41.24	31.90	46.29	37.52	47.88	37.16	39.41	14.38	52.82	42.31	36.50	42.99
gpt-3.5-turbo (few-shot)	48.26	51.93	42.90	57.67	43.41	54.10	45.84	45.29	44.41	64.26	32.35	47.00	50.98
Gold Evidence													
gpt-3.5-turbo (few-shot)	53.40	57.51	50.88	69.15	50.97	62.10	51.05	49.56	56.77	75.35	28.00	53.37	60.47

Table 9: Manual evaluation of decomposed questions. C: Completeness, QU : Question usefulness, EU : Evidence Usefulness. We use the Likert scale of 1-5 and report Cohen's kappa (κ) for inter-annotator agreement.

Method	C (κ)	QU (κ)	EU (κ)
Program-FC	4.6 ±0.77 (0.65)	3.4 ±1.15 (0.53)	2.9 ±1.74 (0.66)
ClaimDecomp	4.5 ±0.86 (0.70)	**3.7 ±0.92** (0.59)	**3.2 ±1.41** (0.69)

Table 10: Automated evaluation of decomposed questions.

Method	Relevance	Diversity
Program-FC	0.782	0.430
ClaimDecomp	**0.831**	**0.490**

ClaimDecomp sees gains of 8.84% in macro-F1 and 10.9% in weighted-F1. This improvement is attributed to more effective evidence retrieval for partially correct claims, as supported by categorical performance. Using original claims sometimes leads to incomplete or null evidence sets. Numerical claim verification requires multiple reasoning steps, as seen in Example 1 from Table 11. Claim decomposition creates a stepwise reasoning path by generating questions on various aspects of the claim, thereby providing necessary information for verification.

5.5 Effect of Different NLI models

To assess the impact of different NLI models, we utilize ClaimDecomp, the top-performing claim decomposition method from Table 7. Table 8 addresses **RQ3**, showing that models trained on numerical understanding, like NumT5-small and FinQA-Roberta-Large, surpass those trained only on general language tasks. Specifically, NumT5-small beats T5-small by 11.8% in macro F1, and FinQA-Roberta-Large, a number-focused Roberta-Large model, exceeds the standard Roberta-Large model by the same margin. The highest performance is achieved by FinQA-Roberta-Large, which also outperforms Roberta-large-MNLI.

Finally, we address **RQ4** by studying the model scale's impact on claim verification reveals that larger models improve performance when fine-tuned, but not necessarily in few-shot or zero-shot settings. For example, gpt-3.5-turbo under-performs in few-shot and zero-shot scenarios compared to smaller fine-tuned models (355M or 60M parameters). This under-performance, observed in flan-t5-xl, gpt-4 and gpt-3.5-turbo is often due to hallucination, where models incorrectly interpret evidence or reach wrong conclusions about claim veracity despite parsing accurate information.

5.6 Performance across different categories of numerical claims

We assessed our fact-checking pipeline's limitations by evaluating baselines in the four categories detailed in Section 4.3. Table 7 shows that methods like ClaimDecomp, which use claim decomposition, outperform original claim-based verification in all categories. Specifically, for comparison based claims ClaimDecomp sees gains of 34.7% in weighted F1 and 31.6% in macro F1 over original claim verification. This is particularly effective for comparison and interval claims, where decomposition aids in handling claims requiring quantity comparisons or reasoning over value ranges, resulting in better evidence retrieval.

In our analysis of different NLI models, fine-tuned models show better performance across all four categories with increased scale. Notably, models with a focus on number understanding, like NumT5-small and FinQA-Roberta-Large, outperform those trained only on language tasks. This is especially relevant for statistical claims that often require step-by-step lookup and numerical reasoning, where FinQA-Roberta-Large achieves a weighted F-1 of **61.36**. Although decomposition approaches and number understanding NLI models enhance performance, explicit numerical reasoning is key for further improvements, a topic for future exploration. A

Table 11: Qualitative analysis of results from different claim decomposition approaches

Method	Decomposition and Verdict
Claim	**Discretionary spending has increased over 20-some percent in two years if you don't include the stimulus. If you put in the stimulus, it's over 80 percent**
Original Claim	**[Verdict]: True**
ClaimDecomp	**[Decomposition]:** [Q1]:Has discretionary spending increased in the past two years?,[Q2]:Does the increase in discretionary spending exclude the stimulus? [Q3]: Is there evidence to support the claim that … **[Verdict]: Conflicting**
Program-FC	**[Decomposition]:** fact_1 = Verify(Discretionary spending has increased over 20-some…), fact_2 = Verify("If you don't include …, discretionary spending has increased…"), fact_3 = Verify("If you put in the stimulus, discretionary spending…"), **[Verdict]: True**
Claim	**Under GOP plan, U.S. families making $86k see avg tax increase of $794.**
Original Claim	**[Verdict]: Conflicting**
ClaimDecomp	**[Decomposition]:** [Q1]:is the tax increase under the gop plan in the range of $794 … making about $86,000?,[Q2]:does the gop plan result in an average tax increase…$86,000?[Q3]:is there evidence that…? **[Verdict]: False**
Program-FC	**[Decomposition]:** fact_1 = Verify("Under GOP plan, U.S. families making $86k…") **[Verdict]: Conflicting**

Table 12: Example of In-context learning sample for GPT-3.5-TURBO few shot for claim verification

Prompts for few-shot fact verification with GPT-3.5-TURBO. Note , the ICL samples are selected dynamically for each test sample:

System prompt: Following given examples, For the given claim and evidence fact-check the claim using the evidence , generate justification and output the label in the end. Classify the claim by predicting the Label: in the end as one of: SUPPORTS, REFUTES or CONFLICTING.
User Prompt:
[Claim]:"A family of four can make up to $88,000 a year and still get a subsidy for health insurance" under the new federal health care law.
[Questions]:is there a subsidy for health insurance under the new federal health care law? can a family of four with an income of $88,000 a year qualify for a subsidy for health insurance? does the new federal health care law provide subsidies for families with an income of $88,000 a year?
[Evidences]:by the way, before this law, before obamacare, …make for example, a family of four earning $80,000 per year ...
Label:SUPPORTS
…Following given examples, for the given claim, given questions and evidence use information from them to fact-check the claim and also additionally paying attention to highlighted numerical spans in claim and evidence. Input […]

detailed ablation study with macro and weighted F1 scores for all categories of numerical claims for different NLI models is shown in Table 8.

5.7 Error Analysis

We conduct an analysis of claims in the test set and their corresponding predictions, offering insights into the considered fact-checking pipeline. Examining results in Table 7 and Table 8, we note the challenge in verifying claims categorized as "conflicting." These claims pose difficulty as they are partially incorrect, requiring the retrieval of contrasting evidence for different aspects of the original claim. We also observe that NLI models with numerical understanding, coupled with claim decomposition, yield better performance. However, there is room for further improvement, as the highest score for this class is only **47.33**.

Among other categories, we observe comparison based numerical claims to be the hard as they are mostly compositional and require decomposition around quantities of the claim followed by reasoning over the different quantities. While claim decomposition helps advance the performance by a significant margin of 31.6% (macro F-1) (Table 7), there are few errors in the decomposition pipeline for approaches like Program-FC. For instance, claim decomposition may result in **over-specification** where the claim is decomposed to minute granularity or **under-specification** where the claim is not decomposed sufficiently. An example of over-specification is shown in the first example for Program-FC in Table 11 where the claim is over decomposed leading to an erroneous prediction. The second example demonstrates a case of under-specification where the claim is not decomposed, leading to limited information and erroneous verdict.

6 CONCLUSIONS

We introduce QuanTemp, the largest real-world fact-checking dataset to date, featuring *numerical data* from global fact-checking sites. Our baseline system for numerical fact-checking, informed by information retrieval and fact-checking best practices, reveals that claim decomposition, models pre fine-tuned using MNLI data, and models specialized in numerical understanding enhance performance for numerical claims. We show that QuanTemp is a challenging dataset for a variety of existing fine-tuned and prompting-based baselines.

A ACKNOWLEDGEMENTS

This work is in part funded by the Research Council of Norway project EXPLAIN (grant number 337133). We also thank Factiverse AI for providing the claim dataset.

REFERENCES

[1] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 85–90. https://doi.org/10.18653/v1/W18-5513

[2] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. arXiv:2106.05707 [cs.CL]

[3] Rami Aly and Andreas Vlachos. 2022. Natural Logic-guided Autoregressive Multi-hop Document Retrieval for Fact Verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6123–6135. https://doi.org/10.18653/v1/2022.emnlp-main.411

[4] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. 2011. Temporal index sharding for space-time efficiency in archive search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 545–554.

[5] Avishek Anand, Srikanta J. Bedathur, Klaus Berberich, and Ralf Schenkel. 2010. Efficient temporal keyword search over versioned text. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, Jimmy X. Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An (Eds.). ACM, 699–708. https://doi.org/10.1145/1871437.1871528

[6] Avishek Anand, Srikanta J. Bedathur, Klaus Berberich, and Ralf Schenkel. 2012. Index maintenance for time-travel text search. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (Eds.). ACM, 235–244. https://doi.org/10.1145/2348283.2348318

[7] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4685–4697. https://doi.org/10.18653/v1/D19-1475

[8] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. BRENDA: Browser Extension for Fake News Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020-07-25) *(SIGIR)*. 2117–2120. arXiv:2005.13270 [cs]

[9] Yixuan Cao, Hongwei Li, Ping Luo, and Jiaquan Yao. 2018. Towards Automatic Numerical Cross-Checking: Extracting Formulas from Text. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1795–1804. https://doi.org/10.1145/3178876.3186166

[10] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. arXiv:2205.06938 [cs.CL]

[11] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. arXiv:1909.02164 [cs.CL]

[12] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (Eds.). Association for Computational Linguistics, Vancouver, Canada, 69–76. https://doi.org/10.18653/v1/S17-2006

[13] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. arXiv:2012.00614 [cs.CL]

[14] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7147–7161. https://doi.org/10.18653/v1/2020.emnlp-main.580

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

[16] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1803–1812.

[17] Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal Archive Search Based on Tags. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 207–210. https://doi.org/10.1145/2872518.2890555

[18] Pegah Jandaghi and Jay Pujara. 2023. Identifying Quantifiably Verifiable Statements from Text. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik (Eds.). Association for Computational Linguistics, Toronto, ON, Canada, 14–22. https://doi.org/10.18653/v1/2023.matching-1.2

[19] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. arXiv:2011.03088 [cs.CL]

[20] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. arXiv:2303.01432 [cs.CL]

[21] Nattiya Kanhabua, Roi Blanco, and Kjetil Nørvåg. 2015. Temporal Information Retrieval. *Foundations and Trends in Information Retrieval* 9, 2 (2015), 91–208. https://doi.org/10.1561/1500000043

[22] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. arXiv:2010.09926 [cs.CL]

[23] Timothy Liu and De Wen Soh. 2022. Towards Better Characterization of Paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8592–8601. https://doi.org/10.18653/v1/2022.acl-long.588

[24] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7787–7813. https://doi.org/10.18653/v1/2023.emnlp-main.483

[25] Tanushree Mitra and Eric Gilbert. 2021. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 258–267. https://doi.org/10.1609/icwsm.v9i1.14625

[26] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4551–4558. https://doi.org/10.24963/ijcai.2021/619 Survey Track.

[27] Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-Hop Fact Checking of Political Claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3892–3898. https://doi.org/10.24963/ijcai.2021/536 Main Track.

[28] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. arXiv:2305.12744 [cs.CL]

[29] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 2173–2178. https://doi.org/10.1145/2983323.2983661

[30] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 22–32.

[31] Anku Rani, S. M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering. arXiv:2305.04329 [cs.CL]

[32] Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 349–361. https://doi.org/10.18653/v1/K19-1033

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[34] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[35] Namika Sagara. 2009. *Consumer understanding and use of numeric information in product claims*. University of Oregon.

[36] Rishiraj Saha Roy and Avishek Anand. 2020. Question Answering over Curated and Open Web Sources. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2432–2435.

[37] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[38] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated Fact-Checking of Claims from Wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 6874–6882. https://aclanthology.org/2020.lrec-1.849

[39] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 65128–65167. https://proceedings.neurips.cc/paper_files/paper/2023/file/cd86a30526cd1aff61d6f89f107634e4-Paper-Datasets_and_Benchmarks.pdf

[40] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 624–643. https://doi.org/10.18653/v1/2021.naacl-main.52

[41] Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. 2016. History by Diversity: Helping Historians search News Archives. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 183–192. https://doi.org/10.1145/2854946.2854959

[42] Nicola Stokes. 2006. Book Review: TREC: Experiment and Evaluation in Information Retrieval, edited by Ellen M. Voorhees and Donna K. Harman. *Computational Linguistics* 32, 4 (2006). https://aclanthology.org/J06-4008

[43] James Thorne and Andreas Vlachos. 2017. An Extensible Framework for Verification of Numerical Claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, 37–40. https://aclanthology.org/E17-3010

[44] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3346–3359. https://aclanthology.org/C18-1283

[45] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[46] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. arXiv:2209.11055 [cs.CL]

[47] Andreas Vlachos and Sebastian Riedel. 2015. Identification and Verification of Simple Claims about Statistical Properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2596–2601. https://doi.org/10.18653/v1/D15-1312

[48] Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. arXiv:2305.16859 [cs.CL]

[49] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[50] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[51] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 485, 13 pages.

[52] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[54] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. arXiv:2203.12990 [cs.CL]

[55] Peng-Jian Yang, Ying Ting Chen, Yuechan Chen, and Daniel Cer. 2021. NT5?! Training T5 to Perform Numerical Reasoning. arXiv:2104.07307 [cs.CL]

[56] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated Fact-Checking: A Survey. arXiv:2109.11427 [cs.CL]

[57] Jiaxin Zhang and Yashar Moshfeghi. 2022. ELASTIC: Numerical Reasoning with Adaptive Symbolic Compiler. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 12647–12661. https://proceedings.neurips.cc/paper_files/paper/2022/file/522ef98b1e52f5918e5abc868651175d-Paper-Conference.pdf

[58] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr