NumTemp: A real-world benchmark to verify claims with statistical and temporal expressions

Venktesh V[1] Abhijit Anand[2] Avishek Anand[1] Vinay Setty[3]

[1]Delft University of Technology, Netherlands [2] L3S Research Institute, Germany
[3]University of Stavanger, Norway

v.viswanathan-1@tudelft.nl aanand@l3s.de avishek.anand@tudelft.nl
vsetty@acm.org

Abstract

Automated fact checking has gained immense interest to tackle the growing misinformation in the digital era. Existing systems primarily focus on synthetic claims on Wikipedia, and noteworthy progress has also been made on real-world claims. In this work, we release **NumTemp**, a diverse, multi-domain dataset focused exclusively on numerical claims, encompassing temporal, statistical and diverse aspects with fine-grained metadata and an evidence collection without leakage. This addresses the challenge of verifying real-world numerical claims, which are complex and often lack precise information, not addressed by existing works that mainly focus on synthetic claims. We evaluate and quantify the limitations of existing solutions for the task of verifying numerical claims. We also evaluate claim decomposition based methods, numerical understanding based models and our best baselines achieves a macro-F1 of 58.32. This demonstrates that NumTemp serves as a challenging evaluation set for numerical claim verification.

1 Introduction

Online misinformation, particularly during elections, poses a significant threat to democracy by inciting socio-political and economic turmoil (Vosoughi et al., 2018). 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 (Guo et al., 2022), driven by the proliferation of open datasets (Thorne et al., 2018; Popat et al., 2018; Chen et al., 2020; Augenstein et al., 2019; Schlichtkrull et al., 2023).

Example: Claim from NumTemp

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 NumTemp

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'(Sagara, 2009), 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 (Hassan et al., 2017) reveals that a substantial 36% of all check-worthy claims in U.S. presidential debates involve numerical quantities or temporal expressions.

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 (Aly et al., 2021). 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 (Chen et al., 2022), MULTIFC (Augenstein et al., 2019) and AVERITEC (Schlichtkrull et al., 2023), 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 (Thorne and Vlachos, 2017; Vlachos and Riedel, 2015), 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 (Fan et al., 2020; Chen et al., 2022; Rani et al., 2023; Aly and Vlachos, 2022; Nakov et al., 2021). 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 **443,320 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 scale 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.
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.

Dataset	# of	Claims	Retrieved	Numerical	†Unleaked
	claims	Source	Evidence	Focus*	Evidence†
Synthetic Claims					
StatProps (Thorne and Vlachos, 2017)	4,225	KB (Freebase)	KB (Freebase)	✓	N/A
Fever (Thorne et al., 2018)	185,445	Wikipedia	Wikipedia	✗	N/A
Hover (Jiang et al., 2020)	26,171	Wikipedia	Wikipedia	✗	N/A
TabFact (Chen et al., 2020)	92,283	Wikipedia	WikiTables	✗	N/A
Feverous (Aly et al., 2021)	87,026	Wikipedia	WikiTables,Wikipedia	✗	N/A
Fact-checker Claims					
LIAR (Wang, 2017)	12,836	Politifact	✗	✗	N/A
ClaimDecomp (Chen et al., 2022)	1250	Politifact	✗	✗	✗
DeClarE (Popat et al., 2018)	13,525	fact-check sites (4)	Web	✗	✓
MultiFC (Augenstein et al., 2019)	36,534	fact-check sites (26)	Web	✗	✗
QABriefs (Fan et al., 2020)	8,784	fact-check sites (50)	Web	✗	✗
AVeriTeC (Schlichtkrull et al., 2023)	4,568	fact-check sites (50)	Web	✗	✗
NumTemp (this paper)	15,514	fact-check sites (45)	Web	✓	✓

Table 1: Comparison of NumTemp 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.

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://anonymous.4open.science/r/NumTemp-E9C0`.

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 (Guo et al., 2022; Botnevik et al., 2020; Zeng et al., 2021). 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 (Zeng et al., 2021; Thorne and Vlachos, 2018). 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 (Aly et al., 2021).

A significant portion of the existing datasets collect claims authored by crowd-workers from passages in Wikipedia (Jiang et al., 2020; Thorne et al., 2018; Aly et al., 2021; Popat et al., 2016; Schuster et al., 2021). Additionally, there are synthetic datasets that require tabular data to verify the claims (Chen et al., 2020; Aly et al., 2021), but these claims and tables may not contain numerical quantities. Recent efforts by (Kamoi et al., 2023) aim to create more realistic claims from Wikipedia by identifying cited statements, but these do not reflect the typical distribution of 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 (Chen et al., 2022; Wang, 2017; Alhindi et al., 2018; Ostrowski et al., 2021), science (Wadden et al., 2020; Vladika and Matthes, 2023; Wright et al., 2022), health (Kotonya and Toni, 2020) and climate (Diggelmann et al., 2021) and other natural claims occurring in social media posts (Mitra and Gilbert, 2021; Derczynski et al., 2017). Multi-domain claim collections like MultiFC (Augenstein et al., 2019) 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, (Vlachos and Riedel, 2015; Thorne and Vlachos, 2017), 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, (Cao et al., 2018)

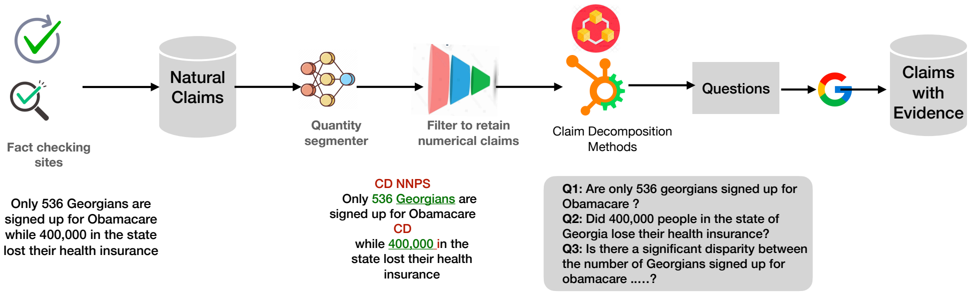
Fact checking sites
Natural Claims
Quantity segmenter
Filter to retain numerical claims
Claim Decomposition Methods
Questions
Claims with Evidence
Only 536 Georgians are signed up for Obamacare while 400,000 in the state lost their health insurance
CD NNPS
Only 536 Georgians are signed up for Obamacare
CD
while 400,000 in the state lost their health insurance
Q1: Are only 536 georgians signed up for Obamacare ?
Q2: Did 400,000 people in the state of Georgia lose their health insurance?
Q3: Is there a significant disparity between the number of Georgians signed up for obamacare .....?

Figure 2: NUMTEMP Construction Pipeline

explore the extraction of formulae for checking numerical consistency in financial statements by also relying on Wikidata. Further, (Jandaghi and Pujara, 2023) 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.

Early fact checking systems simply used the claim as the query to search engine (Guo et al., 2022; Popat et al., 2018; Augenstein et al., 2019) which may not work well if the claims are not already fact-checked. In this regard, recent works have introduce claim decomposition into questions (Chen et al., 2022; Fan et al., 2020; Schlichtkrull et al., 2023; Pan et al., 2023). In this paper, we evaluate the effectiveness of CLAIMDECOMP (Chen et al., 2022) and PROGRAMFC (Pan et al., 2023) 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 (Guo et al., 2022). We use BM25 (Robertson et al., 2009) 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 occurring claims from fact-checking organizations via Google Fact Check Tool APIs[1]. After filtering non-English fact-checkers, it amounts to 45 websites worldwide (listed in Appendix A). Next, we identify quantitative segments (Section 3.2) from the claims and only retain claims that satisfy this criteria. Finally, we collect evidence for the claims (Section 3.4).

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 (Schlichtkrull et al., 2023) (Full mapping is in Appendix A.2). We also ignore those claims with unclear or no labels.

3.2 Identifying Quantitative Segments

We identify quantitative segments in the claim sentence for extracting numerical claims, as defined in (Ravichander et al., 2019), which include numbers, 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 2 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-

[1] `https://toolbox.google.com/factcheck/apis` available under the CC-BY-4.0 license.

assessment of 1000 sample claims from the dataset indicates a 95% accuracy rate for this process.

3.3 Dataset Statistics

After deduplication, our dataset has 15514 claims, divided into training, validation, and test sets with 9935, 3084, and 2495 claims, respectively. The distribution of 'True', 'False', and 'Conflicting' claims is 18.79%, 57.93%, and 23.27%. The dataset is unbalanced, favoring refuted claims, reflecting the tendency of fact-checkers to focus on false information (Schlichtkrull et al., 2023).

3.4 Collecting Evidence

For evidence collection to support claim verification, we adopt the pooling method from Information Retrieval (IR), usually involving top-k document collection from various sources (Stokes, 2006; Santhanam et al., 2022). We modify this method by using the original claim and various decomposition approaches to create a diverse evidence set for veracity prediction retrieval.

We submit original claims to the 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 have evidence from a diverse set of domains including Wikipedia, government websites etc. An overview of top evidence domains in Appendix A.

We enhance evidence diversity by using LLM-based claim decomposition approaches like Program-FC (Pan et al., 2023) and ClaimDecomp (Chen et al., 2022) to generate varied questions and use them as queries. For each generated question, we aggregate the top 10 search results. Following the removal of duplicates and noisy documents, this process results in a comprehensive evidence collection comprising 423,320 snippets.

4 Experimental Setup

To evaluate the NumTemp 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[2] model, pre-fine-tuned on the MNLI corpus, for the NLI task. In Section 4.4 we further explore various NLI models' effectiveness on numerical claims

4.1 Claim Decomposition

ClaimDecomp Prompt

Instructions: You are an assistant that given some examples generates yes/no questions decomposing a claim to aid in fact verification[. . .]
Examples: Claim: $\{c\}$ [Question1]: sq1 [. . .]
Test Claim: [Claim], **Decomposition:** [INS]

We posit that the precise nature of the task of verifying numerical information requires retrieval of relevant evidence containing the required quantitative information. Claim decomposition (Pan et al., 2023; Nakov et al., 2021; Chen et al., 2022; Fan et al., 2020) has been found to be effective in extracting important evidence containing background or implied information to verify the claim.

ClaimDecomp (Chen et al., 2022): The authors 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 NumTemp dataset through in-context learning, setting temperature to 0.3, frequency to 0.6 and presence penalties to 0.8.

Program-FC (Pan et al., 2023): 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.

4.2 Veracity Prediction

Once we have decomposed claims, we use evidence retrieval and re-ranking. Then we employ a classifier fine-tuned on NumTemp 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 3.4. 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 (Reimers and Gurevych, 2019), selecting the top 3 snippets for the NLI task. For the "Original Claim" baseline, we use the top 3 evidences using the claim, and for other methods,

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

we use the top-1 evidence per question, ensuring three evidences per claim for fair comparison.

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 (Wolf et al., 2020) 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 Appendix B.

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 used Setfit (Tunstall et al., 2022), a classifier ideal for small sample sizes, for further annotation. Two annotators manually reviewed 250 random claims, with a Cohen's kappa agreement of 0.58, and 199 claims were correctly categorized. These annotations helped refine the classifier and category assignment.

4.4 Veracity Prediction Model Ablations

We explore various NLI models in the following for veracity prediction:

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 random 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. All prompts are detailed in Appendix E.

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 (Zhang and Moshfeghi, 2022), NUMT5-SMALL (Yang et al., 2021). 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 Hardness of Numerical Claim Verification

To address **RQ1**, we experiment with various claim detection approaches on the NUMTEMP dataset, considering both unified evidence and gold evidence. The performance of different approaches is presented in Table 2. *Fine-tune using Gold Evidence:* Shows performance of NLI models fine-tuned on NUMTEMP (numerical only) and NUMTEMP+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 NUMTEMP 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 NUMTEMP'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.

Method	Statistical		Temporal		Interval		Comparison		Per-class F1			NumTemp Full	
	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													
NumTemp 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**
NumTemp + 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

Table 2: Results of different models on NumTemp (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.

Method	Per-class F1			NumTemp Full	
	T-F1	F-F1	C-F1	M-F1	W-F1
Unified Evidence Corpus					
bart-large-mnli	51.23	79.56	39.37	56.71	64.54
Roberta-large	50.58	77.23	35.50	54.43	62.16
T5-small	19.65	77.22	38.02	44.96	56.89
NumT5-small	36.56	78.45	35.76	50.26	60.26
FinQA-Roberta-Large	49.72	77.91	**47.33**	**58.32**	**65.23**
FlanT5 (zero-shot)	36.35	52.56	3.15	30.68	37.64
FlanT5 (few-shot)	33.90	54.73	20.92	36.52	42.67
gpt-3.5-turbo (zero-shot)	37.81	32.57	31.25	33.87	33.25
gpt-3.5-turbo (few-shot)	44.41	64.26	32.35	47.00	50.98
Gold Evidence					
gpt-3.5-turbo (few-shot)	56.77	75.35	28.00	53.37	60.47

Table 3: Ablation results employing different NLI models for ClaimDecomp on NumTemp.

5.2 Effect of Claim Decomposition on Claim Verification

The **RQ2** is answered by Table 2 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, 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 4. Claim decomposition creates a stepwise reasoning path by generating questions on various aspects of the claim, thereby providing necessary information for verification.

5.3 Effect of Different NLI models

To assess the impact of different NLI models, we utilize ClaimDecomp, the top-performing claim decomposition method from Table 2. Table 3 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 underperforms 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 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.

5.4 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 2 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-

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 4: Qualitative analysis of results from different claim decomposition approaches

	Automated Evaluation		Manual Evaluation		
Method	Relevance	Diversity	Completeness (κ)	Question Usefulness (κ)	Evidence Usefullness (κ)
Program-FC	0.782	0.430	4.6 ±0.77 (0.65)	3.4 ±1.15 (0.53)	2.9 ±1.74 (0.66)
ClaimDecomp	**0.831**	**0.490**	4.5 ±0.86 (0.7)	**3.7** ±**0.92** (0.59)	**3.2** ±**1.41** (0.69)

Table 5: Automated and Manual evaluation of decomposed questions. We use the Likert scale of 1-5 and report Cohen's kappa (κ) for inter-annotator agreement.

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. Additional details and error analysis are in Appendix F and Appendix D.

5.5 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 (Zhang et al., 2020) 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 (Liu and Soh, 2022). 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. We ask two computer scientists familiar with the field to annotate them (guidelines detailed in Appendix C) 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 5.

6 Conclusions

We introduce NumTemp, the largest real-world fact-checking dataset to date, featuring *numerical* and *temporal 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 NumTemp is a challenging dataset for a variety of existing fine-tuned and prompting-based baselines.

7 Limitations

Evaluation of evidence retrieval: Even though we collect an evidence corpus of $423,320$ snippets for the claims through Google search engine, we fix the retrieval stage to use BM25 first-stage retrieval and BERT-based re-ranking. However, it is not necessary that this is the most optimal setup for the fact-checking. There is a need for more thorough evaluation of the evidence retrieval step in the context of downstream fact-checking. In addition, use of better snippet selection methods could also improve the results.

Temporal leakage: While we extensively filter all results from fact-checking websites, we do not consider the temporal leakage avoidance proposed by (Schlichtkrull et al., 2023). There might be still be some web search results which are from non fact-checking domains and contain ready answer to the claim which are published after the claim was published. We estimate this is a minority of the cases. Moreover, as admitted by (Schlichtkrull et al., 2023), their temporal leakage avoidance is also not perfect since, claims are collected from multiple fact-checkers, some may have published the evidence earlier. In addition, from our experience, not all web pages have correct publication dates, which makes the exact dating of the evidence documents challenging.

Numerical Reasoning: Since we deal with numerical claims, which tend to be complex in nature, there is a need for more complex numerical reasoning for the NLI step. In this work, we fine-tune and use models like NUMT5-SMALL (Yang et al., 2021) and FINQA-ROBERTA-LARGE (Zhang and Moshfeghi, 2022), which are pre-trained on numerical understanding data and related tasks. However, these models do not directly perform numerical reasoning. Using the outcome of the numerical reasoning from the evidence provided by answer snippets of multiple questions for predicting the veracity could be a more effective way of fact-checking numerical claims. This approach deserves further research.

8 Ethical Considerations and Risks

Credibility and bias of evidence sources: Fact-checking is a sensitive process, while automated fact-checking systems assist in making some steps of the fact-checking process efficient, they are far from being fully automated. One critical challenge is the credibility of sources of evidence snippets. In this work, we use Google search engine to retrieve evidence which might introduce bias and trustworthiness issues. We predict the veracity of a claim purely based on the textual content of the evidence, but omit the credibility and trustworthiness of the source. Therefore, careful consideration is required before using this work as part of real-world fact-checking processes.

Limitations of LLMs: We employ LLMs for claim decomposition part of the pipeline and as a baseline for the NLI step, which has the risk of hallucination leading to incorrect information. While we control for hallucinations through by grounding the generation on demonstration samples, the effect might still persist.

Additionally, we do not use any private information for the proposed approach. Though some LLMs may have been pre-trained on sensitive information, our prompts do not elicit any sensitive information directly or indirectly.

References

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)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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.

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*, pages 6123–6135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

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*, SIGIR, pages 2117–2120.

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*, WWW '18, page 1795–1804, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims.

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.

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)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims.

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)*, pages 7147–7161, Online. Association for Computational Linguistics.

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, 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*, pages 1803–1812.

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)*, pages 14–22, Toronto, ON, Canada. Association for Computational Linguistics.

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.

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

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims.

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)*, pages 8592–8601, Dublin, Ireland. Association for Computational Linguistics.

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):258–267.

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*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

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*, CIKM '16, page 2173–2178, New York, NY, USA. Association for Computing Machinery.

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*, pages 22–32.

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.

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)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

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*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *arXiv preprint arXiv:2305.13117*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence.

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).

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*, pages 37–40, Valencia, Spain. Association for Computational Linguistics.

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*, pages 3346–3359, Santa Fe, New Mexico, USA. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

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*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches.

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

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

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)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

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.

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.

Peng-Jian Yang, Ying Ting Chen, Yuechan Chen, and Daniel Cer. 2021. Nt5?! training t5 to perform numerical reasoning.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey.

Jiaxin Zhang and Yashar Moshfeghi. 2022. Elastic: Numerical reasoning with adaptive symbolic compiler.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Detailed claim collection process

Claim Statistics We collect natural claims from diverse fact-checking websites. An overview of top-10 commonly occurring websites in our corpus is shown in Table 6. We observe that a significant portion of our claims are from Politifact. We also observe a good proportion of claims from diverse fact checking sources. We also observe from Table 7 that the claims we collect are also from diverse geographical regions.

Claim Source	#Occurences
Politifact	3840
Snopes	1648
AfP	412
Africacheck	410
Fullfact	349
Factly	330
Boomlive_in	318
Logically	276
Reuters	235
Lead Stories	223

Table 6: Top-10 Claim sources and their proportion in our dataset

Country	#Occurences
USA	6215
India	1356
UK	596
France	503
South Africa	410
Germany	124
Philippines	103
Australia	65
Ukraine	35
Nigeria	17

Table 7: Top-10 Claim sources and their proportion in our dataset

A.1 Evidence domains

Table 8 shows the top frequently occurring domains across different categories in our evidence collection. We observe our collection does not have any snippets from manual or automated fact-checkers and related websites or social media handles. Also, the government websites are one of the frequently occurring domains in our evidence collection, as our claims comprise diverse political and international events.

A.2 Label mapping

We collect natural claims from diverse fact checking sites. The label mapping scheme is shown in Table 9. We ignore claims with ambiguous labels like

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

Table 8: Some frequently occurring domains category-wise evidence collection

“Other". In the end, our dataset comprises claims from three categories namely “True", “False" and “Conflicting".

B Numerical claim category distribution

We also categorize claims to the defined categories as explained in Section 4.3. Several examples from each category and their proportion in the whole dataset is shown in Table 10.

C Annotation Guidelines

C.1 Manual Evaluation of Question Quality

We rate the questions generated on 2 aspects: completeness and usefulness (only based on the claim given and top search results). The 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.

Usefulness: This determines if the questions would help verify the claim. One should also consider implicit aspects of the claim when rating this. some questions might be relevant but may just retrieve background knowledge and not relevant to the core aspect being fact checked. use evidence.

While evaluating usefulness, do not make assumptions about verification method. Only check if the questions cover all aspects of the claim and can retrieve relevant evidence and coverage of implied meaning of the claim.

Evidence Usefulness: The annotators are requested to rate the individual 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. Rate the usefulness of all evidences by aggregating usefulness of evidence tied to individual questions. Rate individual evidence based on their relatedness to the claim and their utility in fact checking.

D 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 2 and Table 3, 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 2), 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 4 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.

E Prompt for veracity prediction

The few-shot prompt employed for veracity prediction through GPT-3.5-TURBO is shown in Table 11. We dynamically select few shot examples for every

Original label	**Standard label**
true, correct, fair call, verified, accurate, really true, correct attribution, geppetto checkmark, technically correct, fair, fact, legit, notizia vera, vaccines work	True
false, false/misleading, incorrect, digitally altered, false and missing context, false comparison, altered video, altered photo, altered, mostly false, false claim, false news, unfounded, photo fake, baseless, pants fire, pants-fire, false and misleading, fake, manipulated, lie, manipulation, unproven, not proven, can’t be proven, not provable, wrong, falso, this claim is false, no evidence, just in case, satire, verdadeiro, mas, unsupported, inaccurate, flawed reasoning, really false, labeled satire, originated as satire, scam, bottomless pinocchio, one pinocchio, two pinocchios, three pinocchios, four pinocchios, misleading, not accurate, fabricated, digitally altered, false/missing context, misleading and missing context, conspiracy, public health fakes, pants on fire, not legit, fake news, fiction, totally false, disinformation in the german media, altered photo / video, sarcasm, immagine modificata, altered video/photo, partly fale, miisleading, full-flop, misleading and false, unsubstantiated, trolling, fake tweet, unverified, old video, no proof, no-flip, false connection, recall, fabricated news, fabricated content, misleading content, imposter content, false context / false, fake quote, bad math, bad science, inchequeable, did not happen, wrong numbers, death hoax, the picture in question is morphed, misleading content/partly false, verdict:false, no arrest, not supported, misplaced, fabricated/false, not related, not a cure, recycled hoax, fake letter, staged skit, satirical site, mislaeding, photo out of context, made-up story, recycled rumor, not connected, fake quotes, totally fake, edited video, satirical, false connection/partly false, false/fabricated, fabricated news/ false content, false context/ false, in dispute, hoax!	False
mostly true, mostly-true, missing context, partly false, partly-false, half-true, barely true, barely-true, mixture, mixed, mostly correct, mostly-correct, downplayed, understated, 50/50, cherrypicking, not the full story, overblown, overstated, half true, altered context, altered photo/video, altered image, altered media, mostly accurate, lacks context, partially correct, rather false, partially true, partially false, exaggerated, ambiguous, flip-flop, false headline, partly true, flipflop, inconclusive, miscaptioned, misattributed, two pinocchios, three pinocchios, depends on how you do the math, possibly correct, probably exaggerated, experts are skeptical, flip flop, spins the facts, flip- flop, not quite, mislabeled, misrepresented, mostly not legit, mostly legit, mixed bag, subestimado, correct but, very unlikely, old footage, it’s a joke, the picture is old, not the same, distortion, edited, misleading, not connected, edited video, ambiguous, flip flop, mislabeled, misrepresented	Conflicting

Table 9: Fact-check label mapping from various domains

test example. The examples shown are for a single instance and are not indicative of the examples used for inference over all test examples. We also show the prompt employed for zero-shot veracity prediction in Table 12

F Per-Category results for NLI Ablations

We tabulate the macro and weighted F1 scores for all 4 categories of numerical claims for different NLI models in Table 13.

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%)

Table 10: A broad overview of different categories of claims in NUMTEMP

Prompts for few-shot fact verification with GPT-3.5-TURBO. Note , the ICL samples are selected dynamically for each test sample. Following is an examples and not static for all test samples:

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 strictly one of the following options: SUPPORTS, REFUTES or CONFLICTING.
User Prompt:
[Claim]: Every family health insurance policy has "a $900 hidden tax" to subsidize health care costs of the uninsured.
[Questions]: Is there a hidden tax in every family health insurance policy? Does every family health insurance policy subsidize health care costs of the uninsured? Is the hidden tax in every family health insurance policy $900?
[Evidences]:2009-05-20 in this report families usa quantifies this tax on americans with health insurance coverage. read more. share. the 2010 law says the cost of providing uncompensated care to the uninsured was $43 billion in 2008. ...
Label:SUPPORTS
[Claim]:"Ninety-eight percent of the American people will not see their taxes go up" due to the House health care bill. **[Questions]:** Will 98is there evidence to support the claim that 98is the claim that 98[Evidences]:still, republican leaders in congress and the white house haven't just argued that the bill would have broader economic advantages. . . . ."
Label:CONFLICTING

[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, health insurance rates for much money under the law – these premium increases do make for example, a family of four earning $80,000 per year would save nearly $3,000 per year (or $246 per month) on health insurance premiums. d.h. stamatis. $88,000: new health insurance subsidies will be provided to families of four making up to $88,000 annually or 400% of the federal poverty ...
Label:SUPPORTS

[Claim]:"Starting January 1, 2020, California will tax legal citizens if they don't have health insurance. Why? The state needs to come up with $98,000,000 to pay for free health insurance for illegal aliens."
[Questions]:will california tax legal citizens starting january 1, 2020 if they don't have health insurance? is the purpose of the tax to fund free health insurance for illegal aliens? is the amount needed to fund free health insurance for illegal aliens $98,000,000?
[Evidences]:may 8, 2023 beginning january 1, 2020, california residents must either: have qualifying health insurance coverage; obtain an exemption from the ... illegal aliens only contribute roughly $32 billion in taxes . . . .
Label:CONFLICTING

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 and fact check by thinking step by step and output the label in the end by performing entailment to fact-check claim using the evidence. Classify the entire claim by predicting the Label: as strictly one of the following categories: SUPPORTS, REFUTES or CONFLICTING. Input
[. . .]

Table 11: Example of In-context learning sample for GPT-3.5-TURBO few shot for claim verification

Prompts for zero-shot fact verification with GPT-3.5-TURBO.

System prompt: You are an expert fact checker. For the given claim, questions and evidences fact check the claim. In the end generate the justification first and then the Label: as strictly one of the labels SUPPORTS, REFUTES or CONFLICTING
User Prompt: For the given claim, questions and evidence fact check the claim, generate justification and output the label in the end. Classify the claim by predicting the Label: as strictly one of the following: SUPPORTS, REFUTES or CONFLICTING. Test claim: [Claim], Justification, Verdict:

Table 12: GPT-3.5-TURBO zero shot for claim verification

Method	Statistical		Temporal		Interval		Comparison		Per-class F1			NumTemp Full	
	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	T-F1	F-F1	T/F-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-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 13: Ablation results employing different NLI models for ClaimDecomp on NumTemp. The best results are in **bold**.