TSVer: A Benchmark for Fact Verification Against Time-Series Evidence

Marek Strong and Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{ms2518,av308}@cam.ac.uk

Abstract

Reasoning over temporal and numerical data, such as time series, is a crucial aspect of fact-checking. While many systems have recently been developed to handle this form of evidence, their evaluation remains limited by existing datasets, which often lack structured evidence, provide insufficient justifications for verdicts, or rely on synthetic claims. In this paper, we introduce TSVER, a new benchmark dataset for fact verification focusing on temporal and numerical reasoning with time-series evidence. TSVER contains 287 real-world claims sourced from 38 fact-checking organizations and a curated database of 400 time series covering diverse domains. Each claim is annotated with time frames across all pertinent time series, along with a verdict and justifications reflecting how the evidence is used to reach the verdict. Using an LLM-assisted multi-step annotation process, we improve the quality of our annotations and achieve an inter-annotator agreement of $\kappa = 0.745$ on verdicts. We also develop a baseline for verifying claims against time-series evidence and show that even the state-of-the-art reasoning models like *Gemini-2.5-Pro* are challenged by time series, achieving a 63.37 accuracy score on verdicts and an Ev^2R score of 48.63 on verdict justifications.

1 Introduction

With the growing use of social media and generative AI, there has been an unprecedented increase in the amount of inaccurate and misleading information (Adams et al., 2023; Arnold, 2020). In response, automated fact-checking systems have advanced substantially with the application of new large language models (LLMs) and the development of comprehensive datasets (Vykopal et al., 2024). These systems have shown promising results in identifying and verifying claims across diverse domains and languages (Strong et al., 2024). However, they continue to face challenges when assessing claims that rely on external evidence (Fontana et al., 2025) or when evaluating claims requires deeper reasoning beyond surface-level textual cues (Choi and Ferrara, 2023; Dziri et al., 2023).

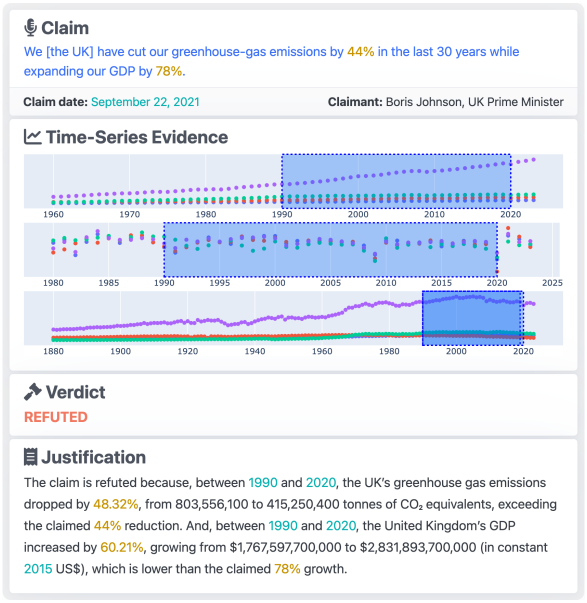

Figure 1: **Example claim from TSVer.** Our dataset includes real-world claims paired with historical time-series evidence. All claims are annotated with time ranges (blue boxes), verdicts, and justifications emphasizing numerical and temporal reasoning.

One of the areas where the reasoning limitations of LLMs are particularly prominent is numerical and temporal reasoning (Akhtar et al., 2023a; Bubeck et al., 2023). This is particularly problematic in the context of fact-checking, where numerical and temporal expressions are prevalent. Studies have shown that over one-third of check-worthy claims involve numerical data (V et al., 2024; Hassan et al., 2017), and many claims demand complex numerical reasoning to be properly evaluated (Aly et al., 2021). Furthermore, as facts can evolve and change over time, fact-checking systems must be

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29906–29926
November 4-9, 2025

capable of accurately interpreting and reasoning over temporal aspects of both claims and supporting evidence (Allein et al., 2023, 2020). Therefore, it is increasingly important to effectively evaluate these capabilities and ensure that fact-checking systems can reliably reason over numerical and temporal data.

In this work, we introduce a new benchmark for evaluating fact verification systems using time-series evidence. Time series is a modality shown to be challenging for language models (Merrill et al., 2024; Fons et al., 2024) and is frequently used by human fact-checkers for fact verification (Akhtar et al., 2023b; Alam et al., 2021). To address this gap, we present **TSVER**—the first benchmark dataset for explainable fact verification grounded in time-series evidence. TSVER pairs real-world claims with historical time-series evidence sourced from fact-checking organizations and includes textual justifications for verdicts, allowing for the evaluation of reasoning about evidence.

To construct TSVER, we collected 287 claims from 38 fact-checking organizations, focusing on those involving numerical and temporal expressions resolvable via time series data. These claims were then aligned with our curated database of 400 time series, extracted from Our World in Data[1]. We avoided claims solvable by simple look-ups or simple arithmetic operations (common in numerical datasets (Lu et al., 2023)) and instead targeted claims requiring reasoning across multiple countries, time series, and claims containing temporal and numerical ambiguities. While time series can be seen as tabular data, their temporal structure and scale add complexity. Compared to prior datasets, TSVER features much larger time series, averaging around 20,000 records per instance, with some over 217,000 records, posing new challenges for fact verification on high-volume, real-world data.

Figure 1 illustrates an example claim from TSVer with annotations for the evidence, verdict, and justification. To fact-check this claim, a system must identify the relevant time series from our dataset (i.e., Greenhouse Gas Emissions, Gross Domestic Product (GDP), and Annual GDP Growth), determine the relevant time frames (i.e., 1990–2020), reason over all data points within these time frames for relevant countries (i.e., the United Kingdom), and generate a verdict accompanied by a justification. Identifying relevant time frames is particularly challenging in our benchmark, as selecting different date ranges often leads to different verdicts. Politicians frequently exploit this by choosing selective dates to support their claims, a practice known as *cherry-picking* (Asudeh et al., 2020). Additionally, a time series may contain both supporting and contradicting periods. For example, while the Gemini-2.5 Pro reasoning model correctly selects 1990 as the starting year (a common baseline for climate-related claims), it uses the period 1990–2019 rather than 1990–2020 to provide supporting evidence. This contradicts the reasoning of our annotators and the original fact-checking article, which notes that Boris Johnson, who made the claim in 2021, relied on outdated figures that ignore the extraordinary impact of the COVID-19 pandemic.

We also propose a fact-checking pipeline as a baseline to demonstrate the feasibility of the task and to benchmark the performance of the state-of-the-art open-weight and proprietary language models. The *gemini-2.5-pro-preview-03-25* reasoning model (Anil et al., 2023), currently ranked first on ChatBot Arena (Chiang et al., 2024), achieves an accuracy of 63.37 on verdict prediction. Additionally, to evaluate models' reasoning in comparison to human annotators, we use the $\mathrm{Ev}^2\mathrm{R}$ scorer (Akhtar et al., 2024), originally developed for evidence retrieval, and demonstrate its effectiveness in this new context. Furthermore, to specifically evaluate evidence retrieval performance with time series data, we introduce a novel metric—TSCS, which jointly measures the accuracy of both time series selection and temporal coverage.

Our dataset is available under a CC-BY-NC-4.0 license at https://github.com/marekstrong/TSVer.

2 Related Work

We summarize key characteristics of existing fact verification datasets in Table 1. In the following , we compare these datasets to TSVER along three key dimensions: evidence modalities, numerical and temporal focus, and the inclusion of human-written justifications.

Evidence Modalities Early fact-checking datasets, such as FEVER (Thorne et al., 2018), primarily relied on textual evidence to support or refute claims. However, as a substantial portion of factual information is embedded in structured sources (e.g., tables, knowledge bases, time series),

[1] https://ourworldindata.org/

Dataset	Domain	#Labels	Real-world Claims	Numerical Focus	Temporal Focus	Evidence Modality	Justifications
FEVER (Thorne et al., 2018)	Multi	3	✗	✗	✗	Text	✗
TabFact (Chen et al., 2019)	Multi	2	✗	✓	✗	Tables	✗
FEVEROUS (Aly et al., 2021)	Multi	3	✗	✗	✗	Text + Tables	✗
AVERITEC (Schlichtkrull et al., 2023)	Multi	4	✓	✗	✗	Text	✓
SciTab (Lu et al., 2023)	Science	3	✓	✓	✗	Tables	✗
Liar++ (Russo et al., 2023)	Politics	2	✓	✗	✗	Text	✓
T-FEVER (Barik et al., 2024b)	Multi	3	✗	✗	✓	Text	✗
T-FEVEROUS (Barik et al., 2024b)	Multi	3	✗	✗	✓	Text + Tables	✗
ChronoClaims (Barik et al., 2024a)	Multi	3	✗	✗	✓	Text	✗
QuanTemp (V et al., 2024)	Multi	3	✓	✓	✓	Text	✓
FinDVer (Zhao et al., 2024)	Finance	2	✗	✓	✗	Text + Tables	✓
TSVer	Multi	4	✓	✓	✓	Time Series	✓

Table 1: Comparison of TSVer with other fact-checking datasets.

subsequent datasets have expanded to include these modalities as well. FEVEROUS (Aly et al., 2021) extends the FEVER framework by pairing synthetic claims with both textual and tabular evidence. FinDVer (Zhao et al., 2024) focuses on tabular data extracted from financial reports, linking it to relevant claims. SciTab (Lu et al., 2023) compiles real-world claims from scientific literature and supports them with tabular evidence. Compared to the previous datasets, TSVER introduces time series as the primary source of evidence.

Numerical and Temporal Focus Since numerical and temporal expressions are common in fact-checking, recent datasets increasingly focus on these aspects. Several benchmarks target numerical reasoning with structured data: TabFact (Chen et al., 2019) verifies crowd-sourced claims against Wikipedia tables, while SciTab (Lu et al., 2023) uses scientific tables to assess compositional reasoning. Domain-specific datasets like FinDVer (finance) (Zhao et al., 2024) combine text and tables with an emphasis on numerical calculations. In open-domain fact-checking, QuanTemp (V et al., 2024) introduces real-world claims involving numerical comparisons and trends, explicitly incorporating temporal reasoning. It is the only existing dataset that targets both numerical and temporal expressions, and since its claims are also sourced from fact-checking websites, it is the closest to our work. While QuanTemp includes justifications, they are unstructured and exclusively textual. In contrast, TSVER provides high-quality, structured justifications for time-series data, including annotations that explain how specific time frames support a claim. Moreover, time series are central in TSVER, whereas it is only one of several claim categories in QuanTemp. Other temporal datasets such as T-FEVER / T-FEVEROUS (Barik et al., 2024b), and ChronoClaims (Barik et al., 2024a) focus more narrowly on date-sensitive or chronological assertions, often using synthetic augmentation or curated timelines.

Justifications Justifying claim verification decisions is a critical component of journalistic fact-checking, reflecting the broader need for transparency and accountability in the verification process (Guo et al., 2022; Kotonya and Toni, 2020). Warren et al. (2025) recently argued that the inherent complexity of fact-checking requires automated systems to offer justifications that allow fact-checkers to critically evaluate their results. Unfortunately, most of the aforementioned datasets do not provide such rationale. A few recent datasets aim to address this gap: AVeriTeC (Schlichtkrull et al., 2023) adds textual explanations synthesizing evidence, LIAR++ (Russo et al., 2023) includes journalist-written justifications, and FinDVer (Zhao et al., 2024) provides expert-annotated step-by-step reasoning.

To the best of our knowledge, TSVER is the only dataset that provides complex structured evidence, focuses on both numerical and temporal claims, and provides justifications for claims' verdicts as well as retrieved evidence. A detailed comparison of TSVER with existing fact-checking datasets across these dimensions is presented in Table 1.

3 Annotation Process

This chapter describes the construction of the TSVER dataset, detailing the end-to-end pipeline from claim extraction to evidence alignment and

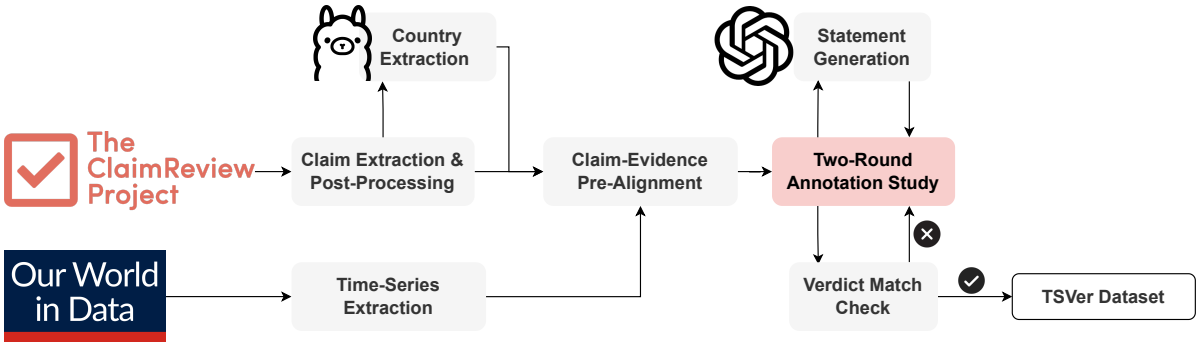

Figure 2: **TSVer data collection pipeline.**

annotation. An overview of the entire process is illustrated in Figure 2.

3.1 Claim Extraction and Post-Processing

To construct our dataset, we began by collecting an initial pool of around 6 thousand claims using The Google FactCheck Claim Search API[2], which aggregates content from the ClaimReview project[3]. We then applied a filtering pipeline to identify claims likely to require reasoning over temporal and numerical data. Specifically, we used *HeidelTime* (Strötgen and Gertz, 2010) to detect the presence of temporal expressions and *spaCy* (Honnibal et al., 2020) to extract numerical expressions from the claim texts. Each retained claim was linked to its corresponding fact-checking article and associated metadata, including the publisher, claimant, and claim date, as provided by the API.

We then post-processed the extracted claims in three steps. First, we manually reviewed cases with missing claim dates and inferred the dates from the accompanying articles. Second, for claims that did not mention any country, we added annotations in square brackets where appropriate (e.g., the annotation '[the UK]' as shown in Figure 1). Lastly, we provided an additional annotation for all countries mentioned in the corresponding fact-checking article. For this step, we prompted the *Llama-3.1-8B* model (Dubey et al., 2024) to identify country names within the raw HTML of each article.

3.2 Time-Series Extraction

As the source of our time series data, we selected Our World in Data (OWID)[4], a non-profit online publication that compiles and curates open-access datasets on global issues such as population, health, economic development, environment, and governance. OWID's focus on socially and politically relevant topics aligns well with the kinds of subjects that are frequently addressed in professional fact-checking. Moreover, OWID is a useful source as it integrates data from a variety of original sources and applies consistent post-processing steps, including the standardization of country names and regional groupings, unit normalization, and computation of derived indicators (e.g., per capita values).

To constrain the complexity of the dataset and simplify alignment with claims, we restricted our scope to time series reported at an annual resolution. However, we also included data sources for which meaningful annual aggregates could be computed from higher-frequency data. In total, we selected 400 time series spanning a wide range of domains. The largest time series in our dataset in terms of the number of records is "Annual Average Surface Temperature," containing 217,899 datapoints. One of the longest in temporal coverage is "Population Density," which extends back to 10,000BC. Many other time series extend back to around 1850, with 1950 being the most common starting year.

Each time series was paired with metadata provided by OWID, including its title, description (including associated notes), and units. Examples of this metadata can be found in Appendix D.

3.3 Claim-Evidence Pre-Alignment

Aligning collected claims with relevant time-series evidence presents some major challenges; in many cases, identifying the correct time series can be as difficult as verifying the claim itself. Time-series data often include semantically related but distinct indicators (e.g., "Cumulative CO2 Emissions", "Per-capita CO2 Emissions", and "CO_2 Emissions"), and for some indicators, data may be available from multiple sources (e.g., "ILO Unem-

[2] https://toolbox.google.com/factcheck/apis
[3] https://www.claimreviewproject.com/
[4] https://ourworldindata.org/

ployment Rate" vs. "IMF Unemployment Rate"). Furthermore, most claims require alignment with multiple time series to be adequately verified (see Figure 1).

Given this complexity, we opted to leave the final alignment of claims with evidence to human annotators (see Section 3.4). However, to reduce annotation effort and ensure consistency, we introduced a pre-alignment step. Specifically, we grouped time series into semantic categories using the OWID taxonomy and matched claims to these groups using keyword-based heuristics. For example, claims containing the term "emissions" were pre-aligned with all time series in the "Environment" category. Claims that could not be aligned with any group were excluded from the dataset.

3.4 Two-Round Annotation Study

We conducted a two-round annotation study using the Prolific platform[5]. Examples of annotation interfaces are shown in Appendix C.

In **Phase 1**, annotators were given a claim, its corresponding fact-checking article, and a set of potentially relevant time series. Their task was to: (i) select relevant time series, (ii) identify all time ranges useful for verification, and (iii) provide explanations for each selected range. Since annotators had access to full articles, they could follow the reasoning of professional fact-checkers, which often included contextual knowledge. For example, fact-checkers often reference years 2005 (Kyoto Protocol) or 2016 (Paris Agreement) in climate-related claims. However, as fact-checkers may have relied on sources beyond OWID, we instructed annotators to prioritize only the provided time-series evidence. To capture contextual information, annotators were also asked to record any useful details from the fact-checking articles (e.g., for resolving ambiguities) that informed their choices of time series and time ranges. These notes will be released as part of the dataset.

In **Phase 2**, new annotators reviewed the evidence annotated in Phase 1 without access to the articles. They were asked to assign one of four verdict labels (see Section 4.1) based on the claim and evidence, and to provide a justification. To assist with reasoning, we presented precomputed statistics (e.g., min/max values, averages, trends) for each time range. The full list is in Appendix F. To avoid overwhelming annotators with irrelevant data, we limited the statistics shown to countries mentioned in the article, as identified during the LLM-based post-processing step (see Section 3.1).

Our initial annotation results showed that while these statistics helped, annotator justifications often lacked numerical expressions and reasoning. To address this, we prompted *gpt-4o-2024-11-20* (OpenAI, 2023) to generate up to five statements based on the provided time-series ranges, focusing on numerical expressions (see Appendix E for prompting details). Annotators were asked to identify truthful ones and optionally incorporate them into their justifications. This modification substantially increased the use of numerical details, resulting in justifications that were more precise and better aligned with the presented time-series data.

As a quality check, we compared annotator verdicts to those from the reference fact-checks. If a majority verdict disagreed with the reference verdict from the article, the claim was re-annotated in a second round. However, we kept the claims afterwards as differences in the evidence may naturally lead to differing verdicts.

3.5 Inter-Annotator Agreement

Following Schlichtkrull et al. (2023) and Ousidhoum et al. (2022), we measured inter-annotator agreement using Randolph's free marginal multirater κ (Randolph, 2005). For verdict labels, we achieve an agreement score of $\kappa = 0.745$, indicating substantial agreement among annotators. For the selection of numerical statements generated by GPT-4o, we observe a lower agreement of $\kappa = 0.581$.

4 TSVER Benchmark

4.1 Dataset Statistics

We collected 287 claims from a total of 38 fact-checking sites. The most represented sites were *Africa Check (25%)*, *Full Fact (13%)*, and *PolitiFact (13%)* (see Appendix D for the full distribution by organization). Following the approach of Schlichtkrull et al. (2023), we adopt a 4-class labeling scheme: *SUPPORTS*, *REFUTES*, *NOT ENOUGH INFO*, and *CONFLICTING EVIDENCE/CHERRY-PICKING*. The dataset is inherently unbalanced, with a higher proportion of *REFUTES* labels (55%). This reflects the nature of fact-checking workflows, where journalists often prioritize addressing false or misleading claims. In terms of geographic coverage, the most claims dis-

[5] https://www.prolific.com/

cuss the United States (29.90%), followed by the United Kingdom (24.23%) and Nigeria (16.49%). A more detailed country-level distribution is provided in Appendix D.

TSVER also includes a curated collection of 400 time series from *Our World in Data*. All time series were preprocessed into a consistent format, and we provide both titles and descriptions to facilitate retrieval. Additionally, inspired by the time-series taxonomy introduced by Fons et al. (2024), we categorize each series by feature type (e.g., trend, volatility, stationarity). To this end, we prompted *Gemini-2.5-Pro* to generate descriptive features for each time series and annotated time ranges.

Few-shot prompting, a technique in which a model is given a few in-context examples, has been shown to improve performance across various fact-checking tasks, including claim detection, evidence retrieval, and general reasoning (Vykopal et al., 2024; Li et al., 2023; Wang et al., 2023). Thus, to support few-shot prompting for our benchmark, we set aside a small development set of 27 claims. To ensure minimal overlap with the test set, these claims are drawn from entirely separate domains, which are explicitly excluded from the test data.

4.2 Synthetic Claims

To further improve the practical utility of this benchmark for training and evaluation, we augmented TSVER by modifying the countries and dates mentioned in claims. We used *gemini-2.5-pro-preview-03-25* to guide this process, generating new claims with different labels. This approach resulted in 300 additional synthetic claims, which will be released as a separate dataset within TSVER. However, this synthetic dataset was not used in our main experiments reported in Section 5.

4.3 Baseline Pipeline

Our baseline pipeline consists of two main components: (1) time-series retrieval and (2) verdict and justification generation.

To retrieve relevant time-series evidence, we rely on the textual metadata (e.g., title, description, units) associated with each time series in the database. Examples of this metadata are provided in Appendix D. While directly using raw time-series data could provide better retrieval performance, exploring methods such as time series encoders (Woo et al., 2024) is beyond the scope of this work.

We prompt an LLM in a few-shot setup, providing examples from the development set, to generate a list of relevant time series as evidence. However, this initial retrieval step often yields too much data: even a few complete time series can exceed the input limits of most LLMs. For example, using Gemini-2.5-Pro, only 31% of the cases had retrieved evidence with fewer than 1 million tokens. Therefore, we apply additional filtering using the same LLM (in a few-shot set-up) to further refine the results. Specifically, we prompt the model to identify relevant time ranges and relevant countries. Note that the model does not have access to fact-checking articles during testing, so it cannot leverage country information in those articles as was done during the annotation phase.

The second baseline component starts by loading the specified slices of time-series data according to the retrieval results. We then prompt the same LLM to generate a verdict along with supporting justifications. Due to the large input size, we adopt a zero-shot setup for this stage. Additionally, for non-reasoning models, we apply Chain-of-Thought prompting (Wei et al., 2022) to explicitly encourage step-by-step reasoning in the output.

All prompting templates are reported in Appendix E.

4.4 Baseline LLMs

Due to the large input sizes resulting from representing time series data in raw text format, our evaluation is restricted to language models with extended context windows. Specifically, we consider only models that support a minimum of 128k tokens. Among proprietary models, we include *Gemini* (Anil et al., 2023) and *GPT* (OpenAI, 2023), while for open-weight models, we choose *Mistral* (Jiang et al., 2023) and *Llama* (Dubey et al., 2024).

For all experiments, the temperature is set to 0.01, top-p to 0.95, and the maximum output length is 4096 tokens.

5 Experiments

5.1 Evaluation Metrics

In addition to standard verdict prediction metrics such as macro-F1 and accuracy, we introduce two complementary evaluation metrics to assess the effectiveness of our retrieval and justification components. These metrics specifically capture the accuracy of time series retrieval with temporal alignment and the factual consistency of generated justi-

fications relative to human justifications.

5.1.1 Time Series Coverage Score

To evaluate the performance of the retrieval component, we assess two key aspects with respect to human-annotated ground truth: (1) whether the correct time series datasets are retrieved, and (2) how well the retrieved time ranges align with the annotated relevant time spans. Both over-retrieval (e.g., retrieving more datasets or time spans than necessary) and under-retrieval (e.g., omitting relevant time series or time spans) can negatively impact downstream performance, either by exceeding the context window or by failing to provide sufficient evidence for verification.

We propose the Time Series Coverage Score (TSCS), a metric that jointly captures the accuracy of both time series selection and temporal coverage. TSCS combines a dataset-level F1 score with a temporal Jaccard Index to evaluate the quality of each retrieval instance.

$$\text{TSCS} = \frac{1}{N}\sum_{i=1}^{N}\left(\text{F1}_i \cdot \overline{J}_i\right) \tag{1}$$

In Equation 1, N denotes the number of evaluation instances. For each instance i, the F1 score is computed, reflecting whether the correct set of time series datasets was retrieved. $\overline{J}_i$ is then the average Jaccard Index over the matched datasets, measuring temporal alignment.

The average Jaccard Index is defined as:

$$\overline{J} = \frac{1}{T}\sum_{j=1}^{T}\frac{|\hat{Y}_j \cap Y_j|}{|\hat{Y}_j \cup Y_j|} \tag{2}$$

Here, T is the number of retrieved time series that correctly match the ground truth data. For each time series j, $\hat{Y}_j$ and Y_j represent the annotated and retrieved time ranges, respectively. The Jaccard Index measures the degree of overlap between these ranges. Averaging across all matched time series yields a robust estimate of temporal accuracy.

5.1.2 Ev^2R Score for Justifications

The Ev^2R scorer (Akhtar et al., 2024) evaluates the quality of evidence retrieval in automated fact-checking by comparing retrieved evidence against reference evidence through atomic fact decomposition. Since the metric essentially compares two free-form texts for factual overlap, we test its suitability for comparing verdict justifications in this context.

The metric comprises three components:

Precision (s_{prec}) This measures the proportion of atomic facts in the retrieved evidence ($A_{\hat{E}}$) that are supported by the reference evidence (E). It is calculated as:

$$s_{\text{prec}} = \frac{1}{|A_{\hat{E}}|}\sum_{a_{\hat{E}} \in A_{\hat{E}}} \mathbb{I}\left[a_{\hat{E}} \text{ supported by } E\right] \tag{3}$$

Recall (s_{recall}) This assesses the proportion of atomic facts in the reference evidence (A_E) that are supported by the retrieved evidence ($\hat{E}$). It is given by:

$$s_{\text{recall}} = \frac{1}{|A_E|}\sum_{a_E \in A_E} \mathbb{I}\Big[a_E \text{ supported by } \hat{E}\Big] \tag{4}$$

F1 Score (s_{F_1}) This is the harmonic mean of precision and recall, providing a balanced measure of the retrieval quality:

$$s_{F_1} = \frac{2\, s_{\text{prec}}\, s_{\text{recall}}}{s_{\text{prec}} + s_{\text{recall}}} \tag{5}$$

In this work, we use the reference-based atomic score from Ev^2R (Akhtar et al., 2024), which was inspired by FactScore (**?**). While we follow a similar prompting template, we adapt it using modified examples drawn from our development set as few-shot instances. We use *gemini-2.5-flash-preview-04-17* as the scorer model, and the prompting details can be seen in Appendix E.

5.2 Results

Our main evaluation results are reported in Table 2. We can observe that even state-of-the-art API models such as *Gemini* and *GPT-4* struggle with the TSVER benchmark, achieving verdict prediction accuracies of 63.37 and 65.35, respectively. This indicates that a large portion of the benchmark remains challenging, even for the most capable commercial models. In contrast, smaller open-weight models, including *Ministral-8B*, *Ministral-3B*, and *Llama-3.1-8B*, perform substantially worse, with accuracies of 38.61, 28.71, and just 7.92, respectively. These results underscore both the difficulty of TSVER and its effectiveness as a probing tool for evaluating model reasoning and verification capabilities.

			Time Series	Verdicts		Justifications		
Model	Params	Max Tokens	TSCS	F1	Accuracy	METEOR	Ev^2R	CL Errors
Gemini-2.5-pro-03-25	-	1M	41.39	68.68	63.37	27.48	48.63	2.31 %
GPT-4.1-2025-04-14	-	1M	33.35	68.58	65.35	31.89	37.57	2.69 %
Mistral-large-2411	123B	128k	26.18	60.11	54.46	32.52	35.75	11.15 %
Ministral-8b-2410	8B	128k	18.04	43.27	38.61	29.66	26.5	9.23 %
Ministral-3b-2410	3B	128k	16.51	36.19	28.71	30.54	21.87	11.92 %
Llama-3.3-70B	70B	128k	26.97	59.32	56.39	27.86	39.52	10.38 %
Llama-3.1-8B	8B	128k	8.21	13.85	7.92	5.88	8.47	78.85 %

Table 2: Verification results with baseline models on the TSVER test set.

When evaluating retrieval quality using the Time Series Coverage Score (TSCS), we observe a substantial performance gap across models. *Gemini* achieves the highest TSCS at 41.39, followed by *GPT-4* with a score of 33.35, indicating that these models are more effective at both selecting the relevant time series and aligning the retrieved time ranges with the human annotated spans. In contrast, smaller models such as *Ministral-8B*, *Ministral-3B*, and *Llama-3.1-8B* perform considerably worse, with TSCS values of 18.04, 16.51, and just 8.21, respectively.

Further analysis of low TSCS scores reveals that smaller models tend to over-retrieve. For instance, while *Gemini* and *GPT-4* retrieve 917 and 871 time series in total across the test set, *Ministral-3B* and *Llama-3.1-8B* retrieve substantially more—3242 and 4893, respectively. This excessive retrieval not only increases the difficulty of downstream reasoning tasks but also places greater demands on context length. As shown in Table 2, smaller models are more likely to exceed their context window limits, leading to context length (CL) inference errors. Notably, *Llama-3.1-8B* failed on 78.85% of the test instances due to exceeding its 128k token limit.

We evaluate justification quality using both METEOR (Banerjee and Lavie, 2005) and the Ev^2R score. While METEOR provides a surface-level measure of lexical overlap with reference justifications, it shows a limited capability to differentiate between our models. For instance, *Ministral-3B* and *GPT-4* obtain comparable METEOR scores of 30.54 and 31.89, despite a substantial gap in their verdict and retrieval performance. This aligns with prior findings of Akhtar et al. (2024), which suggest that surface metrics like METEOR often fail to correlate with human judgments of factual adequacy in explanations.

In contrast, Ev^2R provides a more informative signal by evaluating the factual alignment between model-generated and reference justifications via atomic fact decomposition. According to this metric, *Gemini* leads with an Ev^2R score of 48.63, followed by *GPT-4* at 37.57. All smaller models score lower, with *Ministral-8B*, *Ministral-3B*, and *Llama-3.1-8B* scoring only 26.5, 21.87, and 8.47, respectively. These results suggest that Ev^2R is more sensitive to factual accuracy and evidence relevance in generated justifications, making it a more reliable indicator of model capability in complex verification tasks.

To further probe the complexity of our benchmark, we also conducted experiments with PASTA (Gu et al., 2022), an NLI model specifically designed and pre-trained for numerical and tabular reasoning. PASTA aligns well with TSVER, since it can reason over table-based operations such as column aggregation, min/max comparisons, and row filtering, operations that are commonly required in our dataset. Using the authors' publicly released checkpoint[6], we fine-tuned the model on the TabFact dataset (Chen et al., 2019) and applied PASTA's linearization scripts to convert our time-series tables into a format compatible with the model. Since PASTA only performs binary fact verification, we collapsed all labels other than *SUPPORTED* into the *REFUTED* category. Under this setup, PASTA achieved an F1 score of 43.56, underscoring both the difficulty of our benchmark and the current limitations of table-aware NLI methods when applied to time-series reasoning.

5.3 Discussion

Our baseline system employs a straightforward strategy: formatting raw time series data as Markdown-style tables using `pandas`[7]. While this

[6] https://github.com/ruc-datalab/PASTA

[7] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_markdown.html

representation enables seamless integration with existing LLM-based systems, it presents challenges due to the input length and the continuous, numerical nature of time series data. In particular, the tokenization of floating-point numbers using byte pair encoding (BPE) can yield inconsistent and inefficient representations (Gruver et al., 2023; Spathis and Kawsar, 2023). Furthermore, the sheer length of many time series, often spanning thousands or even millions of data points, can easily exceed the context window limitations of current LLMs.

Unlike traditional fact verification tasks, where input length can often be managed by selecting the top-N most relevant sentences, time series data does not lend itself to such straightforward truncation. Relevant temporal patterns may span long, continuous ranges, making it harder to reduce input size without losing critical evidence.

To address these issues, several studies have introduced quantization-based techniques. For example, models like SpeechGPT (Zhang et al., 2023) and AudioLM (Borsos et al., 2022) employ K-means clustering to convert continuous signals into discrete token sequences, while others use VQ-VAE for a similar discretization process (Duan et al., 2023; Strong et al., 2021). Alternative strategies involve integrating dedicated time series encoders with LLMs, as seen in models such as GPT4TS (Zhou et al., 2023) and Time-LLM (Jin et al., 2023).

Since our results highlight retrieval quality as a critical bottleneck, particularly for smaller models, exploring more efficient time series representations may enable future systems to better encode and reason over temporal data within constrained model budgets.

6 Conclusion

We introduced TSVER, the first benchmark dataset for fact verification grounded in real-world time-series evidence. By focusing on complex claims requiring numerical and temporal reasoning, TSVER show the limitations of current fact-checking systems and large language models in handling structured temporal data. Our LLM-assisted annotation pipeline enables the alignment of claims with time-series evidence, achieving a substantial inter-annotator agreement of $\kappa = 0.745$ on verdicts. The dataset supports rigorous evaluation of both evidence selection and reasoning quality, and we hope TSVER will serve as a valuable resource for advancing research in explainable and evidence-based fact verification.

Limitations

Lack of Multilingual Coverage Although our claims span topics and entities from many parts of the world, we only collected claims and fact-checking articles in English. This design choice simplifies annotation and model evaluation, yet it also means that the benchmark does not assess cross-lingual retrieval, multilingual reasoning, or language-specific numeral and date formats.

Source Bias and Coverage Limitations The claims in TSVER are sourced directly from existing fact-checking articles via the Google Fact Check Explorer. As such, our dataset inherits any biases or limitations present in those original articles. Fact-checkers may differ in how they frame claims, interpret evidence, or articulate justifications, which can introduce variability not reflective of ground-truth facts but rather of editorial choices. Additionally, reliance on a single aggregation tool like the Fact Check Explorer may result in an incomplete or skewed view of the global fact-checking landscape, under-representing claims from less frequently indexed sources or under-covered topics.

Scope of Evidence Our dataset includes only time-series evidence (with textual descriptions), even though real-world claims often require integrating multiple evidence modalities such as reports, tables, charts, or multimedia. While focusing on time series allows us to study a particularly challenging and underexplored aspect of fact verification, the benchmark does not fully represent the broader, multi-modal nature of the fact-checking process.

Ethics Statement

Data provenance and licensing All evidence series in TSVer originate from Our World in Data (OWID), which redistributes underlying statistics from official bodies (e.g., UN, World Bank) under permissive Creative Commons licences (CC-BY 4.0). We preserve the OWID identifiers, metadata, and citations so that the data lineage remains transparent.

Privacy and anonymisation. We did not anonymise any portion of TSVER. All claims are extracted from publicly accessible fact-checking articles that already appear on journalistic websites

and reference well-known public figures, institutions, or countries. These named entities, and the temporal and geographic details, are integral to the factual content of each statement and therefore necessary for fact verification.

Acknowledgement

Marek Strong was supported by the Alan Turing Institute PhD Enrichment Scheme. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958) and the DARPA program SciFy.

References

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463.

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2r: Evaluating evidence retrieval in automated fact-checking. *arXiv*.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023a. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. *arXiv*.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023b. ChartCheck: Explainable fact-checking over real-world chart images. *arXiv*.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *arXiv*.

Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. 2020. Time-aware evidence ranking for fact-checking. *arXiv*, 71:100663.

Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie-Francine Moens. 2023. Implicit temporal reasoning for evidence-based fact-checking. *arXiv*.

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*. FEVEROUS.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Anthropic. 2025. Claude 3.7 sonnet [large language model]. https://www.anthropic.com/claude/sonnet. Accessed: 2025-05-20.

Phoebe Arnold. 2020. The challenges of online fact checking. Technical report, Technical report, Full Fact.

Abolfazl Asudeh, H V Jagadish, You (Will) Wu, and Cong Yu. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment*, 13(6):939–952.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. 2024a. ChronoFact: Timeline-based temporal fact verification. *arXiv*. ChronoClaims.

Anab Maulana Barik, Wynne Hsu, and Mong-Li Lee. 2024b. Time matters: An end-to-end solution for temporal claim verification. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 657–664. T-FEVER. T-FEVEROUS.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. AudioLM: a language modeling approach to audio generation. *arXiv*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. TabFact: A large-scale dataset for table-based fact verification. *arXiv*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv*.

Eun Cheol Choi and Emilio Ferrara. 2023. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. *arXiv*.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2023. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. *arXiv*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *arXiv*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv*.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. 2024. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. *arXiv*.

Nicolo' Fontana, Francesco Corso, Enrico Zuccolotto, and Francesco Pierri. 2025. Evaluating open-source large language models for automated fact-checking. *arXiv*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. *arXiv*.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4971–4983, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv*. Nucleus Sampling.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2023. Time-LLM: Time series forecasting by reprogramming large language models. *arXiv*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. *arXiv*.

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

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. *arXiv*.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. *arXiv*.

OpenAI. 2023. GPT-4 technical report. *arXiv*. GPT-4.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Daniel Russo, Serra Sinem Tekiroglu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *arXiv*.

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

Dimitris Spathis and Fahim Kawsar. 2023. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *arXiv*.

Marek Strong, Rami Aly, and Andreas Vlachos. 2024. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17021–17035, Miami, Florida, USA. Association for Computational Linguistics.

Marek Strong, Jonas Rohnke, Antonio Bonafonte, Mateusz Łajszczak, and Trevor Wood. 2021. Discrete acoustic space for an efficient sampling in neural text-to-speech. *arXiv*.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. 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.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. *arXiv*.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv*.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-checking COVID-19 news claims with scientific evidence. *arXiv*.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. *arXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv*.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. *arXiv*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv*.

Yilun Zhao, Yitao Long, Yuru Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Yiming Zhang, Xiangru Tang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. *arXiv*.

Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One fits all:power general time series analysis by pretrained LM. *arXiv*.

A Experimental Setup

As part of the annotation pipeline, we used the Llama-3 (Dubey et al., 2024) model for inference. Specifically, we employed the 8B-parameter version in 16-bit precision. Inference was performed with a temperature of 1.0 using nucleus sampling (Holtzman et al., 2019), with a top-p value of 0.9.

All annotation scripts and experiments were run on a machine equipped with a single Quadro RTX 8000 GPU (49GB memory) and 64GB of system RAM.

For querying the baseline models, we performed inference using each model's official API. For our baseline experiments with Llama, we used Amazon Bedrock's API instead of the local model to support the full 128k token context window. All API calls were made with default settings unless otherwise specified.

Additionally, we used a combination of GPT-4 (OpenAI, 2023) and Claude (Anthropic, 2025) to assist with parts of the codebase. These models were used as general-purpose coding assistants.

B Annotation details

We carried out our annotations with the help of Prolific (`https://www.prolific.com/`), an online platform which connects researchers with real people willing to participate in studies and surveys, enabling fast collection of high-quality data. The annotations took place on a separate dedicated platform developed by our team and supplied to Prolific.

To ensure high-quality annotations, we applied participant screening criteria available through Prolific. In particular, we restricted participation to individuals located in the United States whose primary language was English and who had completed at least an undergraduate degree (BA/BSc/other). Participants were compensated at an average rate of £10 per hour, in accordance with Prolific's payment principles (`https://researcher-help.prolific.com/en/article/2273bd`).

Figure 3: Instructions given to participants at the beginning of the annotation session. These instructions were followed by a detailed tutorial.

C Annotation Interface

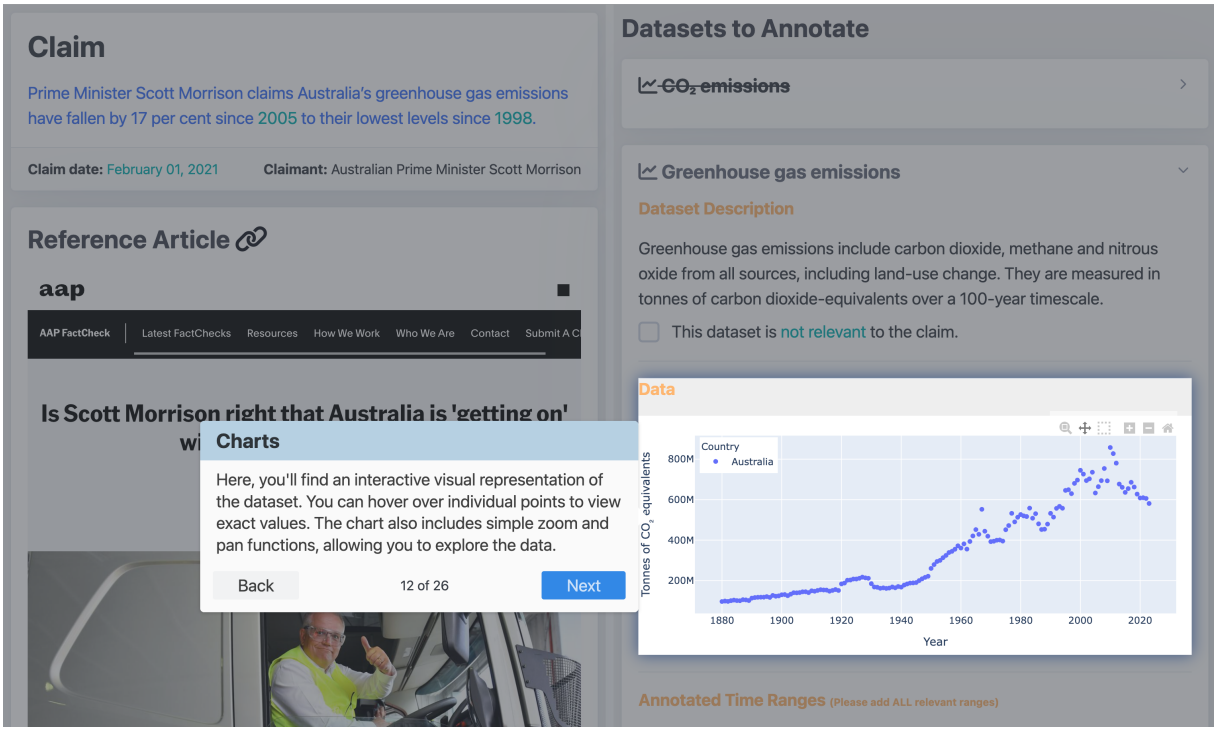

Figure 4: **Detailed step-by-step tutorial explaining the annotation study.**

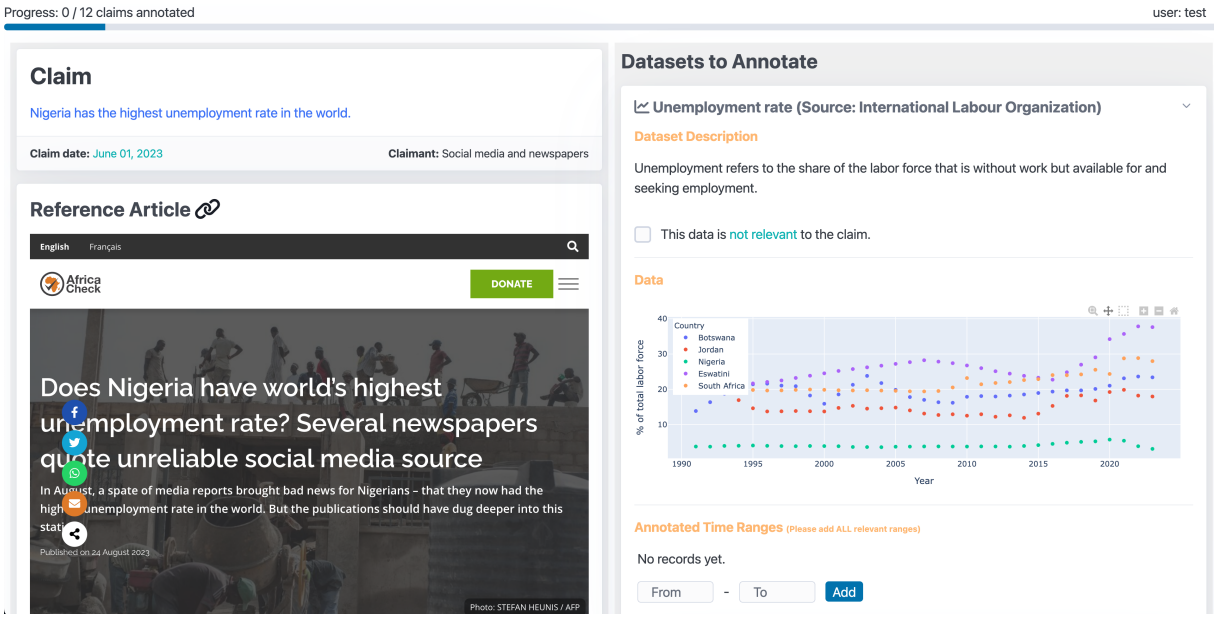

Figure 5: **Annotation Interface for Phase 1.**

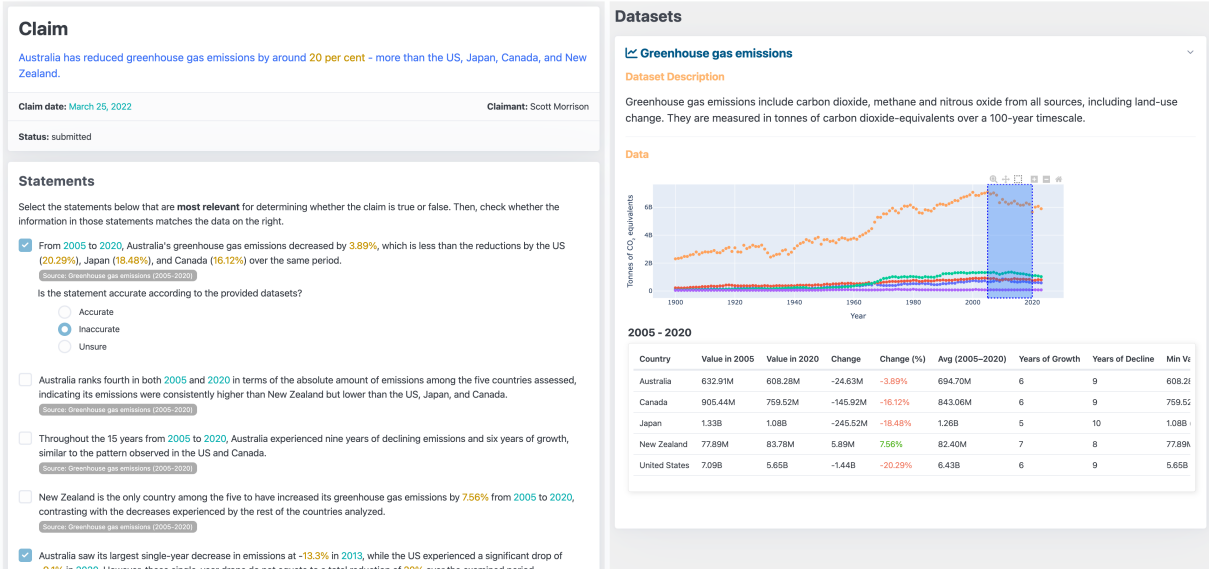

Figure 6: **Annotation Interface for Phase 2.**

D Dataset Details

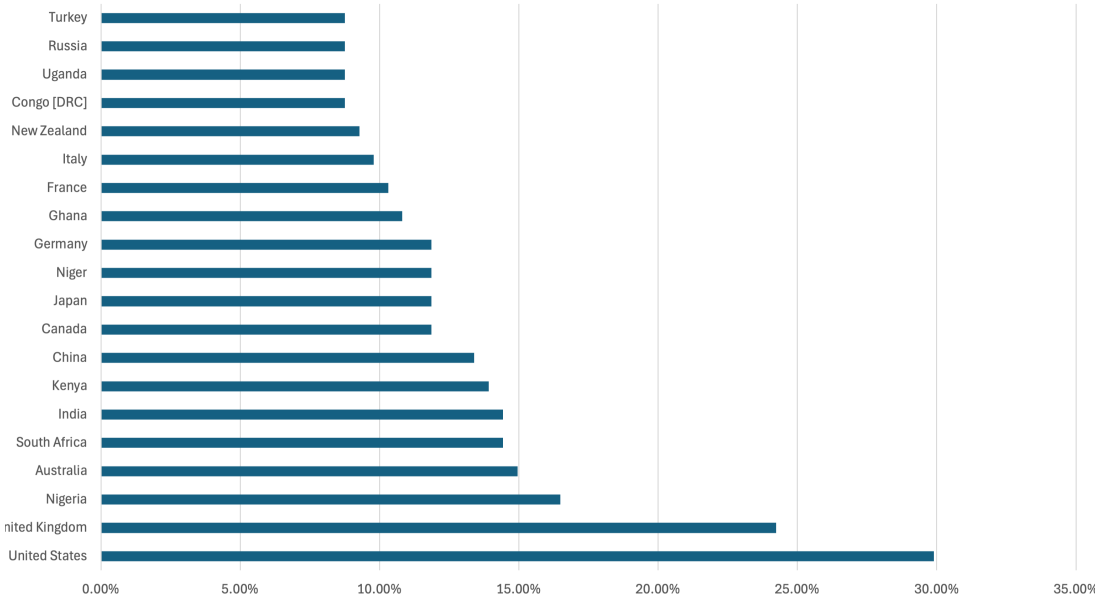

Figure 7: Top 20 countries by share of claims in the benchmark dataset. Bars indicate the percentage of claims associated with each country.

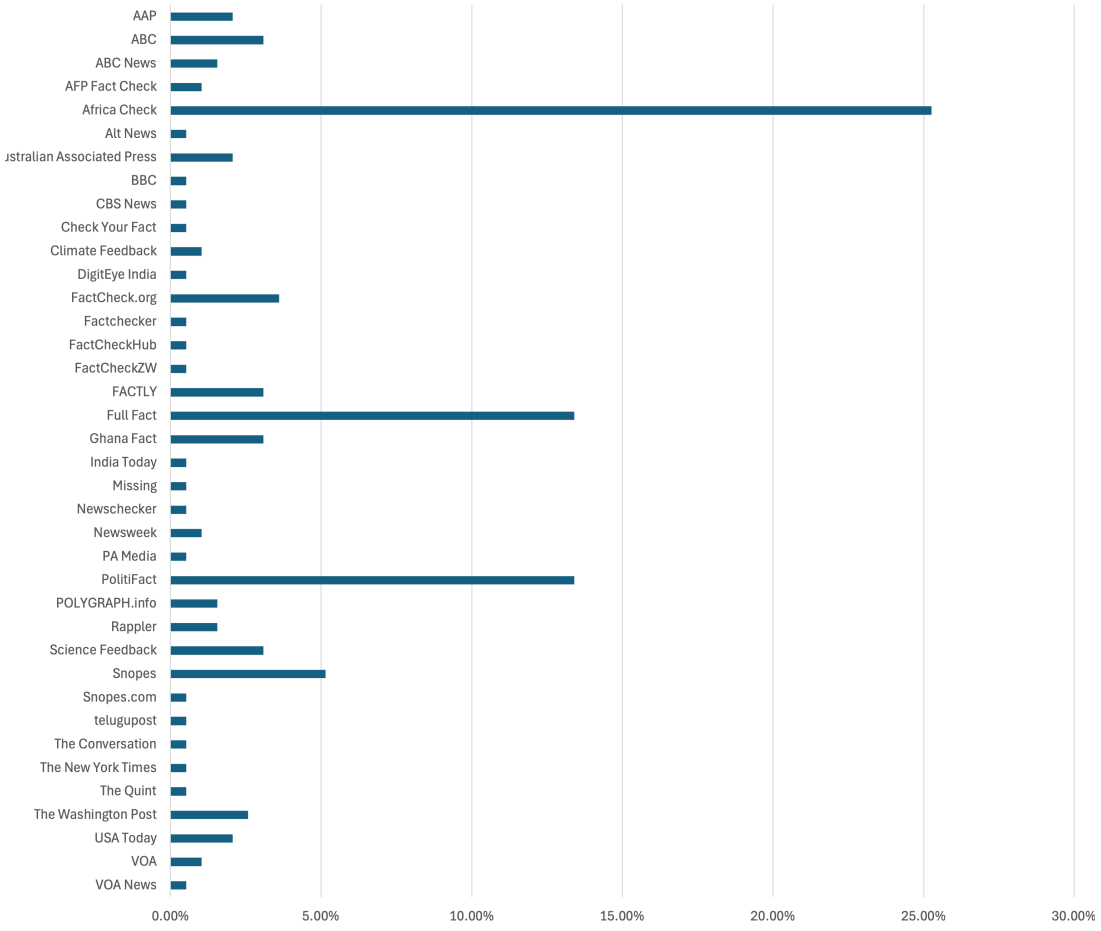

Figure 8: Share of fact-checked claims by publishing organization (*N=38*) in the *TSVer* test set. Africa Check (25%) accounts for the largest share, followed by Full Fact (13%) and PolitiFact (13%).

Name: Population growth rate with migration
Description: "Average exponential rate of growth of the population over a given period. It is calculated as ln(P2/P1) where P1 and P2 are the populations on 1 January of subsequent years."
Unit: "%"

Name: Number of people living in urban areas
Description: "Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.

 Limitations and exceptions: Aggregation of urban and rural population may not add up to total population because of different country coverage. There is no consistent and universally accepted standard for distinguishing urban from rural areas, in part because of the wide variety of situations across countries.

Because the estimates of city and metropolitan area are based on national definitions of what constitutes a city or metropolitan area, cross−country comparisons should be made with caution."
Unit: "People"

Name: Population by age group (15−19 years)
Description: "De facto total population in a country, area or region as of 1 July of the year indicated. This only includes individuals aged 15−19."
Unit: "Persons"

Name: Foreign aid given
Description: "Net official development assistance (ODA) from governments and multilateral organizations, grants from civil society organizations. This data is expressed in US dollars and adjusted for inflation."
Unit: "Constant 2022 US$"

Name: Press freedom index
Description: "The index combines expert estimates with data on violence against journalists. It ranges from 0 (freedom) to 100 (no freedom).

The variable denotes a country's press freedom score. It combines data on violence against journalists with experts assessments by media professionals, lawyers, and sociologists on pluralism, media independence, media environment and self−censorship, legislative framework, transparency, and the quality of the infrastructure that supports the production of news and information."
Unit: "Index"

Name: Per capita CO2 emissions
Description: "
Carbon dioxide (CO2) emissions from fossil fuels and industry. Land−use change is not included. Per capita emissions represent the emissions of an average person in a country or region − they are calculated as the total emissions divided by population.
This data is based on territorial emissions, which do not account for emissions embedded in traded goods.
Emissions from international aviation and shipping are not included in any country or region' s emissions. They are only included in the global total emissions.
"
Unit: "Tonnes per person"

Name: Share of the population that is female
Description: "Female population is the percentage of the population that is female. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship."
Unit: "% of total"

Listing 1: An example of metadata associated with time-series in the TSVer dataset.

E Prompting

```
# Datasets Provided

You have access to the following dataset(s):

{time_series_metadata}

--------------------------

# Data

The following section contains the actual data for each dataset.
Each dataset may include one or more time ranges – when multiple time ranges are available, they are provided separately.

{time_series_data}

--------------------------

# Claim

Evaluate the following claim, stated on {claim_date}:

{claim_text}

--------------------------

# Instructions for Analysis

Based only on the data provided above, generate up to five concise assertions that either support or refute the claim.


Each assertion must follow these guidelines:

1. Data-Backed: Every assertion must cite specific figures, statistics, or rankings from the data. Mention relevant time ranges within the assertion itself. Include the dataset name(s) in brackets at the end of each assertion for attribution, such as [Coal production (2007–2017), Energy consumption (1930–2010), Energy consumption (2003–2004)].

2. Factual Only: Do not include assumptions, interpretations, or projections beyond what the data shows.

3. Data Relevance: Consider whether the data is relevant to when the claim was stated.

4. Clear & Focused: Keep assertions concise and directly tied to the claim.

5. Contextual Language: Refer to data using natural phrasing based on the time periods and content, rather than naming the dataset titles or headings explicitly.

6. Output Format: Use a numbered list (1–5) for your assertions.
```

Listing 2: The prompt template used for statement generation during the second phase of human annotation.

```
# AVAILABLE TIME-SERIES DATA:
{time_series_metadata}
------------------------
# INSTRUCTIONS:
Your task is to identify and list all time-series charts that would be relevant or helpful for verifying a provided claim.
Output your response as a numbered list containing only the titles of the relevant time series charts.
Output the list of relevant time series and say nothing else.
------------------------
# Examples:
{few_shot_examples}
------------------------
# CLAIM:
"{claim_text}"
```

Listing 3: Prompt template for time series retrieval.

```
# TIME-SERIES DATA:
{time_series_metadata}

# AVAILABLE COUNTRIES:
{country_names}
------------------------
# INSTRUCTIONS:
Your task is to identify and list all countries that would be relevant or helpful for verifying a provided claim.
When verifying the claim, we will also have access to time series evidence data as listed above.
Provide your response as a numbered list containing only countries provided in the list above.
Output the list of relevant countries and say nothing else.

------------------------
# Examples:
{few_shot_examples}
------------------------
# CLAIM:
"{claim_text}"
```

Listing 4: Prompt template for the retrieval of relevant countries.

```
# TIME SERIES DATA:

{time_series_metadata}


# CLAIM DATE:

{claim_date}


------------------------


# INSTRUCTIONS:

Your task is to identify all time ranges in the provided time series metadata that could be relevant or helpful for verifying the claim made on {claim_date}.

Output your response as a bullet-point list for each time series, using the following format:

# Time-Series-Name-1
- YYYY-YYYY
- YYYY

# Time-Series-Name-2
- YYYY

# Time-Series-Name-3
- YYYY-YYYY
- YYYY-YYYY
- YYYY-YYYY


Output only the list of relevant time ranges for each time series. Do not include any additional text.


------------------------

# Examples:

{few_shot_examples}

------------------------

# CLAIM:

"{claim_text}"
```

Listing 5: Prompt template for the retrieval of relevant time ranges.

```
Consider the following claim, stated on {claim_date} by {claimant}:
"{claim_text}"

------------------------

Your task is to assess the veracity of this claim using the provided time series data.

# TIME SERIES DATA:

{relevant_tseries_data}

------------------------

# INSTRUCTIONS:
Evaluate the claim in two steps:
– First, select a verdict based on the time series data.
– Second, provide a brief explanation justifying your verdict.

When choosing the verdict, you can choose only from the following options:
{labels_legend}

Format your response as follows:
# VERDICT
...

# EXPLANATION
...
```

Listing 6: Prompt template for verdict and justification generation without CoT.

```
Consider the following claim, stated on {claim_date} by {claimant}:
"{claim_text}"

------------------------

Your task is to assess the veracity of this claim using the provided time series data.

# TIME SERIES DATA:

{relevant_tseries_data}

------------------------

# INSTRUCTIONS:
Evaluate the claim in three steps:
– First, reason about the data step by step.
– Second, select a verdict based on the time series data.
– Third, provide a brief explanation justifying your verdict.

When choosing the verdict, you can choose only from the following options:
{labels_legend}

Format your response as follows:
# REASONING
...

# VERDICT
...

# EXPLANATION
...
```

Listing 7: Prompt template for verdict and justification generation with CoT.

```
You will get as input a claim, a reference evidence and a predicted evidence.
Please verify the correctness of the predicted evidence by comparing it to the reference evidence, following these steps:

1. Break down the PREDICTED evidence in independent facts. Each fact should be a separate sentence.
2. Evaluate each fact individually: is the fact supported by the REFERENCE evidence? Do not use additional sources or background knowledge.
3. Next, break down the REFERENCE evidence in independent facts. Each fact should be a separate sentence.
4. Evaluate each fact individually: is the fact supported by the PREDICTED evidence? Do not use additional sources or background knowledge.
5. Finally summarise (1.) how many predicted facts are supported by the reference evidence, (2.) how many reference facts are supported by the predicted evidence.

Generate the output in form of a JSON as shown in the examples below.

----------

Examples:

{few_shot_examples}

----------

Input:

# CLAIM:
"{claim_text}"

# REFERENCE EVIDENCE:
"{reference_evidence}"

# PREDICTED EVIDENCE:
"{predicted_evidence}"
```

Listing 8: Prompt template for the Ev2R scorer.

F Annotation Details

Name of Operation	Description	Example Output
Difference (Change)	Computes the absolute change in value between the start and end year.	150.0
Percent Change	Computes the percentage change from the starting value to the ending value.	12.5%
Average	Calculates the mean value across the selected years.	123.45
Cumulative Total	Sums all values over the selected time range.	987.65
Standard Deviation	Measures variability or dispersion from the average value.	15.23
Minimum Value	Finds the lowest value in the time range.	100.00
Maximum Value	Finds the highest value in the time range.	200.00
Number of Years of Growth	Counts years where the value increased from the previous year.	4
Number of Years of Decline	Counts years where the value decreased from the previous year.	3
Largest Single-Year Drop	Finds the largest decrease in value between two consecutive years.	-45.67 (in 2012)
Largest Single-Year Increase	Finds the largest increase in value between two consecutive years.	55.23 (in 2018)
Average Rank	Computes the average ranking of a country over the selected years.	2.4
Rank in Year	Ranks countries by value in a specific year (1 = highest value).	1.0

Table 3: List of operators with descriptions used for generating pre-computed statistics in the second annotation phase.