Frame-Guided Synthetic Claim Generation for Automatic Fact-Checking Using High-Volume Tabular Data

Jacob Devasier, Akshith Putta, Qing Wang, Alankrit Moses, Chengkai Li
The University of Texas at Arlington
{jacob.devasier, cli}@uta.edu

Abstract

Automated fact-checking benchmarks have largely ignored the challenge of verifying claims against real-world, high-volume structured data, instead focusing on small, curated tables. We introduce a new large-scale, multilingual dataset to address this critical gap. It contains 78,503 synthetic claims grounded in 434 complex OECD tables, which average over 500K rows each. We propose a novel, frame-guided methodology where algorithms programmatically select significant data points based on six semantic frames to generate realistic claims in English, Chinese, Spanish, and Hindi. Crucially, we demonstrate through knowledge-probing experiments that LLMs have not memorized these facts, forcing systems to perform genuine retrieval and reasoning rather than relying on parameterized knowledge. We provide a baseline SQL-generation system and show that our benchmark is highly challenging. Our analysis identifies evidence retrieval as the primary bottleneck, with models struggling to find the correct data in massive tables. This dataset provides a critical new resource for advancing research on this unsolved, real-world problem.

Keywords: Corpus (Creation, Annotation, etc.), Evaluation Methodologies, Multilinguality, Natural Language Generation, Semantics, (Semi-)Automatic Generation of Training Data, Validation of LRs, Quality Assurance

1. Introduction

Automatic fact-checking has been widely studied in recent years, both as a means of mitigating misinformation and of identifying hallucinations in large language models. Much of this work has focused on using unstructured data, particularly fact-checks from trustworthy sources, both for performing claim matching (Shaar et al., 2020; Putta et al., 2025a) and for training or guiding large language models to generate fact-check verdicts and explanations (Singal et al., 2024; Khaliq et al., 2024; Cheung and Lam, 2023).

Structured tabular data from Wikipedia (Chen et al., 2020; Aly et al., 2021) and scientific documents (Wang et al., 2021; Akhtar et al., 2022) has also been utilized to fact-check claims extracted from Wikipedia (Bouziane et al., 2021) and real-world claims (Wang et al., 2021; Akhtar et al., 2022). However, these studies exhibit two major limitations: (1) they only consider data that has already been processed and prepared for easy consumption by readers and (2) the volume of individual tables is very small, often fitting entirely into the context window of LLMs. These works' reliance on highly-curated data hinders their ability to fact-check novel claims which do not have clean evidence readily available, and their limited volume doesn't fully evaluate the performance of systems at scale.

One previous study (Devasier et al., 2025) explored this direction of using complex, real-world data by collecting factual claims related to OECD (Organization for Economic Co-operation and Development) statistics. [1] However, their dataset was limited in size, containing only ∼70 claims. This small scale is insufficient for comprehensively evaluating a system's retrieval and reasoning capabilities. In this work, we extend this initial concept to create a much larger-scale dataset of synthetic claims for evaluating automatic fact-checking systems on high-volume, complex structured data. We also expand beyond English to include Chinese, Hindi, and Spanish, enabling the evaluation of mul-

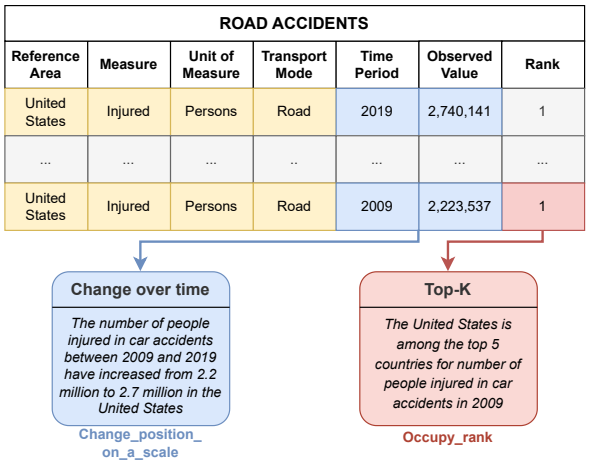

Figure 1: An example of how our system generates factual claims using specific data from a table. Blue and red colors show the data used to create their corresponding claim. Below each claim is the corresponding semantic frame the claim type is grounded in.

[1] https://www.oecd.org/en/data.html

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1721–1731
11–16 May 2026.

tilingual retrieval and reasoning systems.

Our dataset is generated from a large-scale database of 434 high-volume data tables collected by the OECD. To create claims at scale while ensuring their relevance and complexity, we developed a novel generation methodology. We first identified six common claim types inspired by semantic frames (Arslan et al., 2020) frequently occurring in fact-checked claims. For each claim type, we designed algorithms to programmatically select significant data points from the tables. This programmatic selection allowed us to generate a large and diverse dataset. These data points were then provided to an LLM, namely Qwen3-225B-A22B, to generate fluent, natural-language factual claims, as shown in Figure 1.

This process resulted in a dataset of 78,503 synthetic claims, consisting of roughly 47K English claims and ∼10K claims in each of Chinese, Spanish, and Hindi. All claims are mapped to the specific data samples used to create them. To validate the quality of this generation process, we manually reviewed 317 claims and found high quality rates across English (87%), Chinese (91%), and Spanish (88%).

We conducted a novel analysis of the parameterized knowledge within Qwen3 and observed that nearly all specific facts in our dataset are absent from its pre-training corpus. This finding is crucial as a benchmark for this task should measure a system's ability to retrieve and reason over external evidence rather than rely on memorized knowledge. To verify this, we prompted Qwen3 to predict masked factual values directly from claims using only its internal knowledge. The model achieved an exact match in just 2% of cases, confirming that successful performance on our dataset requires genuine evidence retrieval and reasoning, not recall of pre-trained facts.

In addition to the dataset, we implemented a baseline system that employs an LLM to generate SQL queries executable on the OECD database. Our baseline leverages the distinct values of each column to construct these queries, substantially reducing the required LLM context length. Experiments show that our new benchmark is highly challenging. State-of-the-art tabular reasoning systems fail almost completely, and our more robust baseline still struggles. An error analysis reveals the primary bottleneck is evidence retrieval—failing to identify the correct table or extract the correct data—which leads to a high rate of "Not Enough Information" predictions.

Our work differs from prior datasets (Chegini et al., 2025; Zhao et al., 2024; Lu et al., 2023; Aly et al., 2021; Chen et al., 2020) in several key aspects. First, we manually identify types of claims based on the theory of frame semantics instead of letting an LLM have free reign on tables. Second, we algorithmically select specific data samples for each of these claim types, a process guided by our manual definitions that enables scalable generation. Third, and most importantly, our approach generates claims from large-scale tabular data, with each table averaging around 500K rows—a sharp contrast to prior studies that rely on small, highly curated, Wikipedia-like tables. Finally, we generate multilingual claims from the English-only source data.

To summarize, our contributions are as follows:

- We create and release [2] a large-scale, multilingual dataset of 78,503 factual claims mapped to 434 high-volume structured data tables from the OECD.
- We propose a novel, frame-guided methodology for generating claims by programmatically selecting significant and meaningful data samples from complex tables.
- We conduct a novel analysis of parameterized knowledge, demonstrating that our dataset tests active retrieval and reasoning rather than LLM memorization.
- We provide a baseline system and a detailed error analysis, identifying evidence retrieval over high-volume tables as a critical and unsolved challenge for future research.

2. Dataset Generation

Our dataset generation process is designed to create a large-scale, multilingual fact verification dataset grounded in structured OECD data. The process consists of five main stages. First, in *Data Collection*, we compile a large corpus of statistical tables from the OECD (Section 2.1). Second, in *Claim Type Definition*, we define six distinct types of factual claims inspired by common semantic frames evoked in factual assertions (Section 2.2). Third, in *Data Selection*, we develop algorithms to programmatically identify and extract significant data samples from the tables that can form the basis for each claim type (Section 2.3). Fourth, in *True Claim Generation and Curation*, we use a large language model (LLM) to generate natural language claims from these data samples, then employ LLM judges to verify their factual consistency against the source data, producing a clean set of true claims which is subsequently partitioned into train and test splits (Section 2.4). Finally, in *False Claim Generation*, we create a parallel set of false claims by applying systematic, predefined factual perturbations to the true claims (Section 2.5). A high-level overview of this process can be seen in Figure 2.

[2]Dataset can be found at https://github.com/idirlab/megatab-dataset.

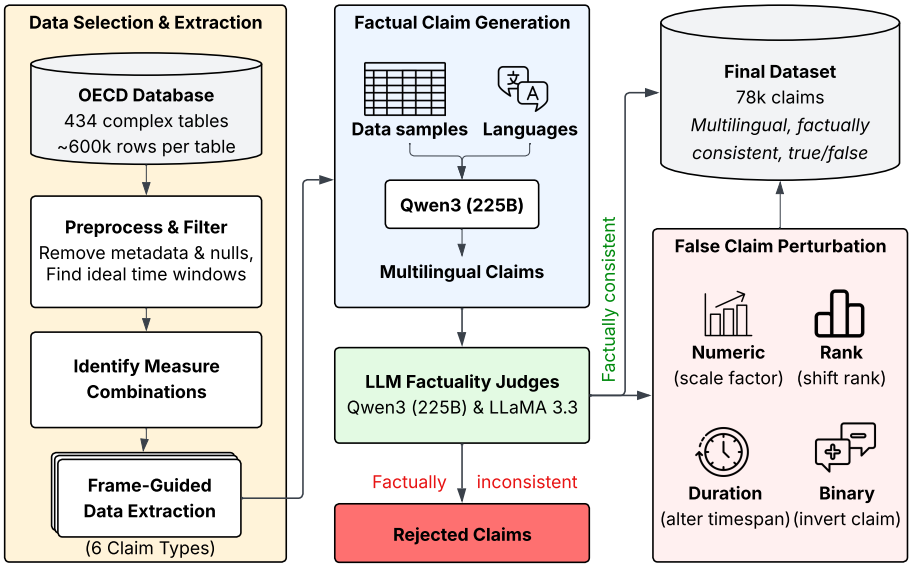

Figure 2: Overview of the dataset generation pipeline. Section 2.2 and Section 2.3 detail the data selection and extraction process, Section 2.4 covers the factual claim generation and LLM judge process, and Section 2.5 explains the false claim perturbation strategies.

2.1. OECD Data Collection

We use the same OECD country statistics as Devasier et al. (2025) for the source data to create our dataset. These statistics cover a wide range of topics related to health, environment and climate change, finance, employment, agriculture, and many others. This data can be explored using the OECD Data Explorer. [3] Collecting these statistics resulted in 434 distinct tables with an average of 596,552 rows per table.

2.2. Claim Type Definition

To generate meaningful claims from the OECD data, we first identified six common types of factual claims. Arslan et al. (2020) analyzed the structure of real-world fact-checked claims and identified a set of recurring semantic frames that underlie them. Drawing on their analysis, we grounded our six claim types in four of these frames: *Change_position_on_a_scale*, *Comparing_at_two_different_points_in_time*, *Comparing_two_entities*, and *Occupy_rank*. We excluded the remaining three frames from their study—*Ratio*, *Recurring_action*, and *Uniqueness_of_trait*—as programmatically identifying interesting and unambiguous data samples for them was particularly challenging. The mapping from frames to our claim types is detailed alongside the data extraction strategies in Section 2.3.

A common concern with fact verification datasets is the complexity of the claims. In previous works, this is typically addressed with either multi-hop information retrieval requirements or factual granularity (e.g., part of the claim is false) (Schlichtkrull et al., 2024b; Khaliq et al., 2024). In this work, we do not generate highly complex claims or multi-claim statements. We justify this approach by noting that most automatic fact verification systems first decompose complex statements into individual atomic claims before verification (Schlichtkrull et al., 2024a; Putta et al., 2025b; Braun et al., 2024; Min et al., 2023).

2.3. Data Selection

For each claim type, we designed algorithms to extract *data samples*—specific sets of rows from the OECD tables that satisfy the criteria for that claim type and can serve as the factual basis for generating a claim.

2.3.1. Data Preprocessing and Filtering

Data structure. Each OECD table contains a collection of columns which we categorize into two primary types: metadata columns and measure columns. Metadata columns contain information shared across nearly all tables, including table names, observation frequencies, and unique identifiers. Measure columns capture the specific measurements conducted within each table. While some measure columns appear frequently across tables (such as measure type, unit of measure, time period, and reference area), others are table-specific (for example, *Transport Mode* in Figure 1).

[3] https://data-explorer.oecd.org/

Many measure columns also have associated identifier columns; for example, a *Reference Area* column with values like “Sweden” might have a related *Reference Area ID* column with values like “SWE”.

Preprocessing. Our data selection process begins with table preprocessing. We first identify and remove metadata columns and measure identifier columns from consideration. We eliminate any columns or rows that contain entirely null values. We remove rows where the primary observation value (the numerical data point found in the *obs_value* column) is null or where the *observation status* is not classified as *normal* (e.g., ‘Estimated’ values for future dates). Tables lacking *Reference Area* (e.g., country) information are also excluded.

Time-window filtering. To ensure claims are generated from time periods with robust and comparable data, we identify an "ideal time window" for each table. We first find the time period with the maximum number of reporting countries (the "maximum coverage"). We then select time periods where the country count is within 5% of this maximum. This "leniency" allows for small, temporary dips in reporting countries (e.g., due to different reporting frequencies like quarterly vs. annual) within an otherwise stable time window. Tables that do not contain at least a two-year ideal time window and include at least 20 countries are excluded.

Measure combinations. Following preprocessing, our cleaned tables typically contain reference area, time period, and a collection of measure columns, along with the primary observation value. To structure the data for claim extraction, we retrieve the unique combinations of non-numeric measure column values. For example, a table might contain three measure columns: *Measure*, *Unit of Measure*, and *Transport Mode*. We identify all occurring unique combinations of these values, such as (`Injury crash`, `Crashes`, `Road`) or (`Fatalities`, `Persons`, `Road`).

Each such *measure combination* defines a specific analytical perspective on the table (e.g., *the number of crashes where someone was injured in a road accident* for (`Injury crash`, `Crashes`, `Road`)). All subsequent data extraction strategies (Section 2.3.2) operate within individual measure combinations to find significant data points, which we refer to as *data samples*.

2.3.2. Claim-Specific Data Extraction

Next, we implement six distinct data extraction strategies, one for each claim type. These strategies are applied within each measure combination to find significant data points.

Top-K claims. These claims identify countries ranking at the extremes for a measure, following the pattern: “Country X is among the top/bottom 5 countries on measure M in year Y.” **Logic:** For each measure combination and year, we rank countries based on their *observed value*. We select the top and bottom k data samples. We set $k = 5$ for subsets with >50 countries (for that measure combination and year) and $k = 3$ for subsets with 20–50 countries. Subsets with <20 countries are excluded.

Constant-Change claims. These claims capture sustained trends, following the pattern: “Country X has shown a constant increase/decrease on measure M for N consecutive years, as of year Y.” **Logic:** We identify continuous spans of at least eight years showing a consistent increase or decrease in the *observed value* for a country within a measure combination. In the claim, N represents this identified span (e.g., $N = 8, 9, ...$).

Historical-Extreme claims. These claims identify locally unprecedented performance, following the pattern: “In year Y, Country X recorded its highest/lowest value on measure M in the last N years.” **Logic:** For each country and measure combination, we analyze its time-series data to find the maximum number of years (N) since a higher (or lower) value was recorded. We require $N \geq 10$ years to ensure the claim represents a notable event.

Change-in-Rank claims. These claims document significant shifts in relative performance, following the pattern: “Country X went from rank A to rank B on measure M between year Y and year Z.” **Logic:** We identify instances where countries experience substantial changes in their relative rankings over time. A change is “substantial” if it is: (1) a flat change of at least 10 positions or 20% of the total number of countries in that subset, whichever is larger; or (2) a change-in-rank ratio of at least 2 for the same country (e.g., moving from rank 10 to rank 5).

Change-Over-Time claims. These claims focus on absolute value changes, following the pattern: “Country X went from value A to value B on measure M between year Y and year Z.” **Logic:** These claims use the same underlying data points as the Change-in-Rank claims but emphasize the change in observed values rather than positional changes in rankings.

Have-Trait claims. These claims provide simple factual statements, following the pattern: “Country X has value A on measure M in year Y.” **Logic:**

Claim Type	English	Chinese	Hindi	Spanish
Change-in-Rank	10,350	2,208	2,207	2,123
Have-Trait	9,089	1,897	1,913	1,882
Change-Over-Time	8,928	1,988	2,004	1,974
Top-K	8,827	2,012	1,969	1,860
Constant-Change	5,093	1,784	1,606	1,410
Historical-Extreme	4,544	908	933	925
Total	46,831	10,797	10,632	10,174

Table 1: Number of samples in our dataset broken down by language and claim type.

We generate these claims using data from the individual temporal data points (e.g., the start and end points) identified for the Change-in-Rank and Change-Over-Time claim types, creating separate atomic claims for each time point.

2.4. Claim Generation and Curation

Our data selection process (Section 2.3) yielded 104,930 data samples that met our extraction criteria (e.g., being a top-k value, a historical extreme, etc.). To avoid biasing our dataset towards tables with especially high volume or claim types which are more abundant, we limited the number of data samples drawn from each OECD table for each claim type, aiming to generate a maximum of 100 claims in English and 20 in each non-English language.

We generated claims for each data sample using Qwen3-225B-A22B. We faced two primary concerns with using LLM-generated outputs to build a dataset. First, the factuality of claims with respect to the intricate and detailed facts of the world are difficult to ensure. Second, it may be difficult to communicate a fact clearly without sometimes losing subtle details. This can lead to or be exacerbated by low linguistic quality of the sentence.

To address the factual consistency of the claims, we used two LLM judges (Llama 3.3 70B and Qwen3-225B-A22B) to verify that the generated claim was factually supported by the provided data sample (detailed in Section 4.2). To address linguistic quality, we relied on Qwen3's thinking mechanism, which allows the model to refine its output rather than generating it in a single pass. In total, our dataset consists of 87,517 claims determined to be factually consistent by both LLM judges.

Data partitioning. To build useful train and test partitions, we ensured our test set would evaluate out-of-domain performance by holding out at least 10% of the OECD tables entirely from the training set. This partitioning process resulted in a final training set of 75,666 claim-data pairs and a test set of 2,837 claim-data pairs (totaling 78,503 claims). The remaining 9,014 claims (from the original 87,517) were set aside to ensure a clean separation between in-domain and out-of-domain tables. We sampled roughly 10% of these test set claims for human evaluation (Section 4.1). The language and claim type counts for the final dataset are shown in Table 1.

2.5. Creating False Claims

Note that the LLM-based curation process described above was designed to ensure the *true* claims are factually consistent; the false claims are then derived from these verified true claims through controlled modifications. We define four types of perturbations based on the claim type: **Numeric** perturbations, where the given value is multiplied by a randomly selected scaling factor (e.g., 0.5, 1.5, 2.0); **Rank** perturbations, where the rank is randomly shifted by a substantial amount (e.g., moving a rank to a much higher/lower position); **Duration** perturbations, where the duration in years is randomly altered, for instance by shifting a start/end year by 2–6 years or extending a duration by 3–8 years; and **Binary** perturbations, where the claim's direction is inverted (e.g., *increased* → *decreased*, or *top-K* → *bottom-K*). While these template-based perturbations produce reliably false claims, we acknowledge that they may not capture the full diversity and subtlety of real-world misinformation; exploring more naturalistic false claim generation is an avenue for future work.

3. Baseline System

We implement a baseline system to extract evidence and perform reasoning over large, structured tables. We store the OECD Data tables in a SQL database, for ease of storage and retrieval. Prior approaches (Chen et al., 2020; Lu et al., 2023) typically serialize entire tables into text sequences provided to a language model within its input context. However, this design quickly becomes infeasible for large tables, as the serialized representation can exceed the context length limits of even modern LLMs. Our baseline instead decomposes the problem into modular retrieval and reasoning steps that avoid such scaling bottlenecks.

Given a claim c, the system first uses an LLM to decompose it into a set of atomic subclaims $c_1, \ldots, c_k$, where the number of subclaims is determined dynamically by the LLM at inference time using in-context learning. This atomic fact decomposition enables the system to retrieve heterogeneous evidence—potentially from different tables or from multiple regions within a single table—that would not be captured by a single SQL query.

For each subclaim, the system retrieves the most semantically similar table using the

gte-multilingual-base embedding model [4] and its reranker.[5] To support this search, we precompute textual representations for all tables in the OECD dataset. Each table representation concatenates the table name, the OECD-provided description, and the names and representative values of categorical data columns. These representations are embedded using the aforementioned models and form the searchable corpus for table retrieval. The table with the highest cosine similarity to a subclaim's embedding is selected for further processing.

Once a relevant table is identified, the system prompts an LLM to generate an executable SQL query that extracts the evidence needed to verify the subclaim. The prompt includes the claim text, table name, OECD description, column names, and a subset of unique categorical values for each column.[6] For columns with more than 20 unique values, we use BM25 retrieval to select the 20 most similar values relative to the subclaim. The SQL generation step is retried up to three times if execution fails, similar to the SQL generation in Devasier et al. (2025).

After retrieving the relevant data, the system prompts the LLM again to determine whether each subclaim is *True*, *False*, or *Not Enough Information (NEI)* based on the query results. Finally, these subclaim-level judgments are synthesized into an overall verdict for the original claim using another LLM call.

4. Experiments

4.1. Human-Annotated Quality Check

To evaluate the quality of the LLM-generated synthetic claims, we manually annotated 317 (10% + additional oversampling for English) claims randomly selected from the test set, including 143 English, 66 Chinese, 59 Spanish, and 49 Hindi claims. Each language was annotated by a native speaker of the respective language. Each annotator was instructed to label the data-claim pair as "good" or "bad" and provide a comment for any bad claims.

We found similar quality rates across English (87%), Chinese (91%), and Spanish (88%) with each being rated "good" around 90% of the time. Hindi, however, had a much lower rate of only 66%. Based on the annotator's comments, we found that 13% of the Hindi samples had only minor issues related to the use of "among OECD countries" within the claim which we later determined to be acceptable, resulting in an actual good quality rate of 79%.

		Llama 3.3		
		True	**False**	**NEI**
Qwen3	**True**	87,517 (84.2%)	579 (0.6%)	4,695 (4.5%)
	False	4,679 (4.5%)	441 (0.4%)	339 (0.3%)
	NEI	1,683 (1.6%)	33 (0.0%)	3914 (3.8%)

Table 2: Distribution of samples predicted to be True, False, or NEI by two LLM-based factuality evaluators.

4.2. LLM-as-a-Judge Factuality Check

We also performed a more robust evaluation (Table 2) of the factuality of the entire dataset using two different LLM judges: Qwen3-225B-A22B and Llama 3.3 70B. We chose to include Llama 3.3 as a judge to mitigate bias that occurs between models within the same family (Li et al., 2025; Panickssery et al., 2025). Both LLMs were prompted to verify the factual correctness of the generated claims based on the data sample and the dataset description. Each claim was classified as True, False, or Not Enough Information (NEI), and was accompanied by a justification generated by the LLM. Llama 3.3 frequently had trouble with exact value mismatches (e.g., it would classify a claim as False if any numeric rounding was involved), so we adjusted the prompt to allow for minor rounding differences based on the LLM's judgement (i.e., by stating the values "should be within rounding error"). The distribution of the judged outputs from this experiment is shown in Table 2.

We found that 16% of the generated claims were not predicted to be True by at least one of the models. Constant-Change claims exhibited the lowest proportion of claims predicted True by both (75%). Upon examining a sample of 100 claims, we observed that this was typically caused by the way the data was presented to the LLM. A similar trend was observed for other claim types with Llama 3.3, where many predictions cited insufficient supporting data, leading to a relatively high overall rate of Not Enough Information (NEI) predictions (8.6%). We also found that most cases in which Qwen3 predicted a claim as False were due to minor rounding differences in numerical values. Such variations were permitted during claim generation to make the claims sound more natural.

4.3. Baseline Performance

In this experiment, we compare our baseline system (Section 3) with TabSQLify (Nahid and Rafiei, 2024), a state-of-the-art system in tabular reasoning, on the human-evaluated subset of the test set. TabSQLify originally used GPT-3.5 with a context

[4] https://huggingface.co/Alibaba-NLP/gte-multilingual-base

[5] https://huggingface.co/Alibaba-NLP/gte-multilingual-reranker-base

[6] The *obs_value* column is excluded from the prompt.

Model	Tables	Accuracy (%)
TabSQLify	Predicted	3.9 ± 1.3
Baseline no-think	Predicted	11.3 ± 2.0
Baseline	Predicted	15.5 ± 0.7
TabSQLify	Predicted-HQ	3.8 ± 1.5
Baseline no-think	Predicted-HQ	14.9 ± 2.6
Baseline	Predicted-HQ	**20.4** ± **2.6**
TabSQLify	Gold	6.4 ± 1.5
Baseline no-think	Gold	26.4 ± 2.3
Baseline	Gold	36.0 ± 0.8
TabSQLify	Gold-HQ	6.4 ± 1.6
Baseline no-think	Gold-HQ	32.9 ± 2.0
Baseline	Gold-HQ	**42.7** ± **5.9**

Table 3: Accuracy of our baseline system on the human evaluation dataset. HQ indicates that we excluded samples marked as bad by humans.

length of 3K tokens, so to ensure a fair comparison we reimplemented their system with Qwen3-30B-A3B with a 15K token context length. This comparison is presented in Table 3.

As we expected, TabSQLify failed to handle the high-volume tables in the OECD dataset, resulting in an accuracy of 6.4% on the gold tables and 3.9% when using the tables predicted by our baseline. Our baseline performed significantly better on both predicted and gold tables, with the thinking-enabled LLM consistently performing the best, with 36% and 20.4% accuracy on the gold and predicted tables, respectively. After removing low-quality ("bad") test samples identified by human annotators (Section 4.1), we observed a significant improvement in our baseline's performance, coupled with a larger variance between runs. However, this improvement was not observed for TabSQLify, suggesting that its failures are unrelated to the quality of the claims.

We also evaluated our system's ability to retrieve the correct evidence from the database, as shown in Table 4. For this experiment, we used the entire test set to obtain more robust metrics. Overall, we found that the `gte-multilingual` semantic similarity models performed well across all language for both evidence retrieval tasks (table retrieval and data retrieval), though the models performed noticeably worse on Hindi table retrieval than the other languages for the True verdict predictions.

As expected, the table retrieval accuracy is highest for verdicts predicted as True. This is also the least predicted verdict, as the LLM seems to only predict True when it is more confident. The primary source of error for our baseline appears to be NEI predictions, which account for 53% of all predictions. False claims appear to pose less of a challenge, as the LLM can predict claims to be false using contradicting evidence that does not necessarily originate from the same data as the claim.

4.4. Parameterized Knowledge Leakage

In this experiment, we aim to analyze the internal knowledge of Qwen3-30B-A3B to understand how much of our baseline's performance can be attributed to the LLM's internal knowledge obtained during its pretraining. Previous studies (Almeida et al., 2025; Mousavi et al., 2024; Moayeri et al., 2024) have also evaluated the knowledge recall abilities of LLMs in this domain; however, the facts they used were very common and therefore more likely to be known by the LLMs. We design two experiments using the Have-Trait claims from our dataset's test set. We choose this claim type due to its simplicity and higher likelihood of the LLM retrieving the corresponding fact.

Masked fact prediction. First, we test whether the specific facts in our dataset are directly represented within the model parameters. Intuitively, if an LLM encodes a fact, it should be able to predict the corresponding value in that fact. To evaluate this, we mask the numerical value of the measure within each claim and prompt the model to predict the missing value.

We measure the deviation between the LLM's predicted value and the ground truth value and define a relative tolerance parameter p that sets the acceptable range of values to be between $[\frac{v}{1+p}, v(1+p)]$, where v is the ground truth value. As shown in Table 5, the LLM is able to predict the masked value exactly ($p = 0$) for roughly 2% of the claims. We also observe a rapid increase in performance between $p = 0$ and $p = 0.5$, indicating that the LLM often approximates the correct answers; however, it is evident that most specific facts are not precisely represented within the model's internal knowledge.

Claim-knowledge alignment. Second, we test whether the model's world knowledge is consistent with the generated claims. To do this, we prompt the LLM with a given claim and directly ask it to predict whether the claim is *True* or *False*. As shown in the confusion matrix in Table 6, the LLM has a heavy bias toward predicting a claim to be *False* (89%) regardless of its truthfulness. The LLM achieves a precision of 61.8%, a recall of 12.7%, and an F1-score of 21.1%.

To determine whether the LLM genuinely possesses the knowledge required to correctly predict the masked values in the claims—rather than arriving at them by chance—we analyze the consistency of its predictions across these two experiments. To do this, we only consider the true claims, as predicting a false claim to be *False* does not necessarily entail the fact is known by the LLM. The results are presented in Table 7.

Predicted Verdict	Overall	English	Spanish	Chinese	Hindi
Baseline table retrieval accuracy					
Overall	32.2%	36.6%	33.5%	29.3%	30.0%
True	75.0%	78.6%	77.9%	77.8%	67.2%
False	36.8%	40.1%	38.0%	33.8%	35.6%
NEI	21.3%	22.7%	24.8%	19.5%	18.6%
Baseline data retrieval accuracy					
Overall	18.3%	22.8%	18.5%	14.9%	17.3%
True	52.2%	52.7%	54.7%	55.2%	48.0%
False	21.0%	24.9%	21.7%	18.6%	19.1%
NEI	10.4%	13.2%	11.8%	6.9%	10.0%

Table 4: Subclaim-level evidence retrieval performance on table retrieval (top block) and data retrieval (bottom block). The Verdict Accuracy column shows the accuracy of the predicted verdicts, duplicated for both retrieval blocks. The overall accuracy refers to the weighted average of all the verdicts.

Tolerance Level	Accuracy (%)	Δ
$p = 0$	2.0	–
$p = 0.25$	14.9	645%
$p = 0.5$	24.4	64%
$p = 1.0$	36.1	48%
$p = 1.5$	42.3	17%
$p = 2.0$	47.1	11%

Table 5: Accuracy of masked fact predictions across different tolerance levels.

		Predicted	
		True	**False**
Label	**True**	996 (6.8%)	6,838 (46.7%)
	False	614 (4.2%)	6,195 (42.3%)

Table 6: Confusion matrix for knowledge-claim alignment.

Correct Predictions		Tolerance Level	
Task 1	Task 2	$p = 0$	$p = 0.1$
✓	✓	0.4%	2.5%
✓	✗	1.6%	5.0%
✗	✓	12.5%	10.4%
✗	✗	85.5%	82.1%

Table 7: Consistency of LLM predictions between masked fact prediction (Task 1) and alignment with generalized knowledge (Task 2) across different tolerance levels.

We found that only 0.4% of the claims in our test set were correctly predicted ($p = 0$) in both of these tasks, meaning that the LLM has likely only been exposed to a few of the claims in our dataset during its pretraining. At $p = 0.1$, this increases to 2.5%. Based on these results, we conclude that nearly all of the performance from our baseline can be attributed to the system itself rather than any internal knowledge held by the LLM.

5. Related Works

Table-to-text generation. Zhao et al. (2023) investigated the table-to-text generation capabilities of large language models (LLMs) in real-world information-seeking scenarios. Their analysis concluded that open-source LLMs still lag behind GPT-4 in generating text from tables; however, the gap between open and closed-source models has closed significantly since their work was published. Many previous works (Zhang et al., 2024; Iravani and Conrad, 2024; Min et al., 2024; Sundararajan et al., 2024; Gong et al., 2019) have studied table-to-text generation. However, these approaches typically grant LLMs excessive freedom in text generation, often producing outputs that are overly general, verbose, or unfocused with respect to a specific fact. In contrast, our work aims to generate concise, fact-specific statements grounded in data. To achieve this, we guide the LLM to produce claims that align with semantic frames commonly observed in real-world fact-checked claims (Arslan et al., 2020).

Fact verification with tabular data. FEVEROUS (Aly et al., 2021) is a fact verification dataset using Wikipedia tables and text as evidence. The claims in FEVEROUS were generated by taking sentences from Wikipedia articles and mutating them to create false claims. TabFact (Chen et al., 2020) (and RePanda (Chegini et al., 2025), an extension of TabFact) is another table-based fact verification dataset using Wikipedia tables. FINDVER (Zhao et al., 2024) is a recent benchmark consisting of long financial documents with many tables in the documents. SCITAB (Lu et al., 2023) is another table-based fact verification dataset focusing on scientific tables. However, all of these

datasets use relatively small tables (less than 100 rows on average) and do not consider high-volume data as we do in this work.

Querying high-volume data with LLMs. Devasier et al. (2025) is the primary motivation for this work. They collected a dataset of real-world factual claims from trustworthy fact-checking sources and used OECD data to fact-check them. However, their dataset was very limited in size (roughly 70 claims) due to the difficulty of collecting real-world claims from the OECD datasets. In this work, we instead generate synthetic claims which can be fact-checked using OECD data, allowing us to create a much larger dataset. Radhakrishnan et al. (2024) also built a system to integrate high-volume data with LLMs to improve factual accuracy of responses to user's statistical queries. They chose to use Data Commons, a large collection of public datasets from trusted organizations, much like the OECD datasets. This work trained an LLM to convert natural language queries into Data Commons queries which were then mapped to a set of templated queries to retrieve relevant data. One limitation of their work is that they do not query their data with SQL, which is a widely studied task with LLMs.

6. Conclusion

In this work, we address a significant gap in automated fact-checking by moving beyond small, curated tables to high-volume, complex structured data. We introduced a novel, frame-guided methodology to generate a large-scale, multilingual dataset of 78,503 synthetic claims. Grounded in massive OECD data tables, our dataset is not only substantial in size but also diverse, covering English, Chinese, Spanish, and Hindi, and is structured around six common semantic frames to ensure claims are realistic and varied.

Our experiments validate the challenge and utility of this new benchmark. We demonstrated through knowledge-probing experiments that the facts in our dataset are not typically memorized by LLMs, meaning systems must perform genuine evidence retrieval and reasoning rather than relying on parameterized knowledge. Our baseline system, which decomposes claims and generates SQL queries, struggled to perform well on our system (9.3% on the full test set). Our error analysis revealed the primary bottleneck is evidence retrieval—specifically, failing to find the correct table or extract the correct data, leading to a high rate of "Not Enough Information" predictions.

Our work has several limitations that suggest directions for future research. First, all claims are synthetically generated by a single LLM family, which may introduce linguistic artifacts or structural regularities not found in naturally occurring claims. Second, false claims are produced through template-based perturbations (e.g., numeric scaling, rank shifting), which may be more predictable than real-world misinformation. Third, our evaluation compared only against TabSQLify; benchmarking against modern agent-style systems or hybrid retrieval-augmented pipelines would better contextualize the difficulty of our benchmark. Fourth, several design choices in the data selection pipeline—such as the top-k thresholds, the 5% time-window leniency, and the eight-year minimum for Constant-Change claims—may influence the distribution of claim types, though we leave a systematic sensitivity analysis to future work. Finally, Hindi claim quality was lower than the other languages, and further investigation into the root causes and targeted improvements for multilingual generation are needed.

Despite these limitations, the difficulty our baseline faced demonstrates that fact-checking on large-scale, complex structured data remains a largely unsolved problem. This dataset provides a valuable new resource for the research community to develop and benchmark more robust models, underscoring the critical need for future work on improved table retrieval and precise data extraction techniques.

7. Bibliographical References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Thales Sales Almeida, Giovana Kerche Bonás, João Guilherme Alves Santos, Hugo Abonizio, and Rodrigo Nogueira. 2025. TiEBe: Tracking language model recall of notable worldwide events through time. *arXiv preprint arXiv:2501.07482*.

Rami Aly, Zhijiang Guo, M. Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, O. Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and verification over unstructured and structured information. *ArXiv*, abs/2106.05707.

Fatma Arslan, Josue Caraballo, Damian Jimenez, and Chengkai Li. 2020. Modeling factual claims with semantic frames. In *Proceedings*

of the Twelfth Language Resources and Evaluation Conference, pages 2511–2520, Marseille, France. European Language Resources Association.

Mostafa Bouziane, Hugo Perrin, Amine Sadeq, Thanh Nguyen, Aurélien Cluzeau, and Julien Mardas. 2021. FaBULOUS: Fact-checking based on understanding of language over unstructured and structured information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 31–39, Dominican Republic. Association for Computational Linguistics.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. DEFAME: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*.

Atoosa Chegini, Keivan Rezaei, Hamid Eghbalzadeh, and Soheil Feizi. 2025. RePanda: Pandas-powered tabular verification and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32200–32212, Vienna, Austria. Association for Computational Linguistics.

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.

Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853.

Jacob Devasier, Akshith Reddy Putta, Rishabh Mediratta, and Chengkai Li. 2025. Task-oriented automatic fact-checking with frame-semantics. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13825–13842, Vienna, Austria. Association for Computational Linguistics.

Heng Gong, Xiaocheng Feng, Bing Qin, and Ting Liu. 2019. Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time). 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 3143–3152, Hong Kong, China. Association for Computational Linguistics.

Sahar Iravani and Tim O. F. Conrad. 2024. Towards more effective table-to-text generation: Assessing in-context learning and self-evaluation with open-source models. *ArXiv*, abs/2410.12878.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference Leakage: A contamination problem in llm-as-a-judge. *ArXiv*, abs/2502.01534.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li, and Qianren Wang. 2024. Exploring the impact of table-to-text methods on augmenting LLM-based question answering with domain hybrid data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 464–482, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1211–1228, New York, NY, USA. Association for Computing Machinery.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. DyKnow: Dynamically verifying time-sensitive factual knowledge in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029, Miami, Florida, USA. Association for Computational Linguistics.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024. TabSQLify: Enhancing reasoning capabilities of llms through table decomposition.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2025. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. 2025a. ClaimCheck: Automatic fact-checking of textual claims using web evidence. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 303–316, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. 2025b. ClaimCheck: Real-time fact-checking with small language models.

Prashanth Radhakrishnan, Jennifer Chen, Bo Xu, Prem Ramaswami, Hannah Pho, Adriana Olmos, James Manyika, and R. V. Guha. 2024. Knowing when to ask – bridging large language models and data.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024a. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos, editors. 2024b. *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics, Miami, Florida, USA.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, Miami, Florida, USA. Association for Computational Linguistics.

Barkavi Sundararajan, Yaji Sripada, and Ehud Reiter. 2024. Improving factual accuracy of neural table-to-text output by addressing input problems in ToTTo. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7350–7376, Mexico City, Mexico. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Haowei Zhang, Shengyun Si, Yilun Zhao, Lujing Xie, Zhijian Xu, Lyuhao Chen, Linyong Nan, Pengcheng Wang, Xiangru Tang, and Arman Cohan. 2024. OpenT2T: An open-source toolkit for table-to-text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–269, Miami, Florida, USA. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.