Un*kn*own Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data

Jean-Flavien Bussotti* **Luca Ragazzi*** **Giacomo Frisoni***
Gianluca Moro* **Paolo Papotti***
EURECOM, France `{bussotti, papotti}@eurecom.fr`
Department of Computer Science and Engineering, University of Bologna, Italy
`{l.ragazzi, giacomo.frisoni, gianluca.moro}@unibo.it`

Abstract

Computational fact-checking (FC) relies on supervised models to verify claims based on given evidence, requiring a resource-intensive process to annotate large volumes of training data. We introduce UNOWN, a novel framework that generates training instances for FC systems automatically using both textual and tabular content. UNOWN selects relevant evidence and generates supporting and refuting claims with advanced negation artifacts. Designed to be flexible, UNOWN accommodates various strategies for evidence selection and claim generation, offering unparalleled adaptability. We comprehensively evaluate UNOWN on both text-only and table+text benchmarks, including FEVEROUS, SCIFACT, and MMFC, a new multi-modal FC dataset. Our results prove that UNOWN examples are of comparable quality to expert-labeled data, even enabling models to achieve up to 5% higher accuracy. The code, data, and models are available at `https://github.com/disi-unibo-nlp/unown`

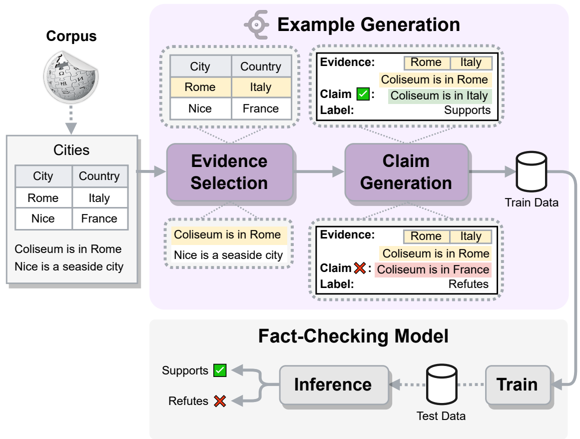

Figure 1: UNOWN pipeline. Given a corpus of documents, the *Example Generation* module (investigated in this work) outputs training instances.

1 Introduction

The spread of false information on social media threatens public trust. For example, during the COVID-19 pandemic, misinformation led to vaccine hesitancy, straining public health systems and informed decision-making (Saakyan et al., 2021; Carey et al., 2022; Carrieri et al., 2023). Computational fact-checking (FC) is a vital tool for verifying claims against diverse evidence types, including unstructured text and structured tabular data. Diversity increases task complexity, requiring advanced NLP methods to cross-reference information accurately (Guo et al., 2022).

Traditional FC verification models (i.e., those making final predictions over evidence, without retrieving it) heavily rely on training samples manually annotated by experts, who meticulously review and pair claims with corresponding evidence, and intentionally modify claims to create refuting examples. Unfortunately, this process is labor-intensive and time-consuming, which significantly hinders the scalability of FC efforts in adapting to evolving misinformation scenarios (Nakov et al., 2021). Recent studies have attempted to mitigate these challenges by automating the generation of training examples using question-answering (QA) and entity replacement (ER) algorithms (Pan et al., 2021; Wright et al., 2022). Yet, as shown in Table 1, they face limitations that restrict their practical utility:

1. *They fail to integrate precise tabular data with nuanced textual data*, which is often essential for verifying real-world claims (Chen et al., 2020; Aly et al., 2021); see Figure 2.
2. *They are confined to specific domains*, such as biomedicine, due to their reliance on vertical knowledge bases (KBs), compromising their ability to generalize across fields.

To overcome these limitations, we present UNOWN (Figure 1),[1] a novel approach that uses pre-

*Equal contribution (co-first authorship).

[1]Pronounced "unknown", the name draws inspiration from the cryptic Pokémon hieroglyphs (i.e.,), reflecting the uncertain factuality label of undisclosed textual claims.

Work	U+S†	Domain Agnostic	Tested Datasets	Human Eval‡
Pan et al. (2021)	✗	✓	FEVER (2018)	✗
Wright et al. (2022)	✗	✗ (biomed.)	SCIFACT (2020)	✓
Ours	✓	✓	FEVEROUS (2021) SCIFACT (2020) MMFC (*new*)	✓

† The study combines unstructured and structured data as evidence.
‡ The study includes human examination of the generated examples.

Table 1: Summary of works on the automatic generation of training samples for fact-checking systems.

trained language models (PLMs) to generate synthetic training examples for FC systems, integrating textual and tabular evidence.[2] Unlike prior work relying on ER methods and domain-specific data, UNOWN offers a flexible solution that supports multiple evidence selection and claim generation strategies, accommodating small and large language models (SLMs and LLMs). This versatility not only broadens the system's utility across real-world applications but also facilitates its deployment in diverse hardware environments, from low-power devices to advanced computing systems.

We validate our approach by comparing the accuracy of FC models trained on examples generated by UNOWN versus those labeled by humans.[3] To achieve this, we conduct extensive experiments on text-only and text+table evidence scenarios using three public FC datasets targeting general and scientific content: FEVEROUS (Aly et al., 2021), SCIFACT (Wadden et al., 2020), and MMFC, our new multi-modal and multi-domain fact-checking dataset.[4] MMFC complements FEVEROUS as the second existing corpus featuring textual and tabular evidence, distinguishing it from SCIFACT, which exclusively focuses on text.

The main findings of our study are as follows:

- In text-only evidence scenarios, training on UNOWN data yields lower accuracy, showing an 8% gap compared to human-labeled samples. However, this gap diminishes to just 2% with the inclusion of only 100 human-labeled instances. Conversely, in text+table scenarios, we achieve up to 5% higher accuracy.
- SLMs and LLMs produce synthetic data of comparable quality, with just a 1% gap in downstream FC accuracy.
- By transcending traditional reliance on external KBs, UNOWN adeptly generates refuting claims with sophisticated negation artifacts.

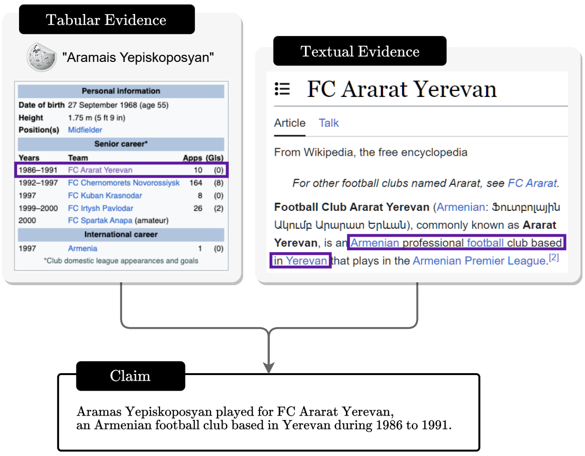

Figure 2: Example from the FEVEROUS dataset where the verification of dates reported in the claim requires reasoning above both textual and tabular information.

2 Related Work

Computational FC has been an active area of research for decades (Dagan et al., 2005; Guo et al., 2022). Recently, the rise of LLMs has advanced the development of FC pipelines (Schulman et al., 2022), but their effectiveness is still inferior to human experts (Saeed et al., 2022; Caramancion, 2023). Specialized models are currently the most effective approach (Li et al., 2023), despite they require large labeled datasets for training.

Existing methods for automatically generating FC training examples have been approached through both unsupervised and supervised techniques. Unsupervised solutions, typically employed in the absence of labeled data, leverage PLMs to create textual claims from a given text, e.g., by using template prompts (Meng et al., 2022). Supervised approaches rely on specific resources, e.g., an annotated taxonomy to train an LSTM model for sentence generation (Meng et al., 2019).

Several works have investigated the generation of claims from textual evidence (see Table 1). Pan et al. (2021) produce question–answer pairs using answer replacement to assemble the refuting claim. Wright et al. (2022) create supporting claims with a generative PLM and ER over a domain-specific KB for evidence refusal in the biomedical field.

Research on generating claims specifically from tabular data remains limited. While some stud-

[2]This work focuses on the FC verification sub-component, excluding evidence retrieval, and thus UNOWN is not intended as a replacement dataset for a complete FC system.

[3]We employ state-of-the-art FC models as they existed at the start of this study (March 2023), without changing their original implementations and hyperparameters.

[4]The dataset is available in the HuggingFace hub: https://huggingface.co/datasets/disi-unibo-nlp/MMFC

ies have explored template-based methods (Wang et al., 2021; Veltri et al., 2023), Bussotti et al. (2023) demonstrated improved results by producing claims based on human-provided examples.

Artificial text passages have recently demonstrated greater effectiveness than human-written ones for reasoning-demanding QA (Frisoni et al., 2024), but FC tasks have not yet been studied.

To the best of our knowledge, no existing work has addressed the generation of FC training examples from structured and unstructured data as input.

3 Problem Formulation

Let $\mathbf{d}$ represent a semi-structured document (e.g., a Wikipedia page) containing n sentences and m tables. We define evidence $\mathbf{e} = \{\mathbf{e}_s, \mathbf{e}_t\}$ as a non-empty subset of sentences $\mathbf{e}_s = \{s_1, \ldots, s_{|\mathbf{e}_s|<n}\}$ and, optionally, cell values $\mathbf{e}_t = \{c_1, \ldots, c_{|\mathbf{e}_t|<p}\}$ extracted from a table within $\mathbf{d}$, where p is the total number of cells. A supervised FC model $\mathcal{F}$ evaluates whether a textual claim c is supported or contradicted by the given evidence $\mathbf{e}$. Specifically, $\mathcal{F}$ takes as input a data pair $<\mathbf{e}, c>$ and outputs a verdict from the set $\mathcal{L} = \{\textit{Supports}, \textit{Refutes}\}$.[5] Consequently, our goal is to automatically generate labeled examples $\mathcal{E} = <\mathbf{e}, c, l \in \mathcal{L}>$ to train $\mathcal{F}$.

Challenge I: Refuting Claims. There have been proposals to generate artificial claims by synthesizing $\mathbf{e}$ in a sentence. Abstractive summarization has been explored with text-only evidence (Tonguz et al., 2021; Wright et al., 2022) and scenarios centered on cell values only (Bussotti et al., 2023). In contrast, our goal is to create claims that incorporate evidence from both structured and unstructured data, as illustrated in Figure 1. However, while a *Supports* claim naturally aligns with the provided evidence, we also require examples with a *Refutes* label to train FC models effectively, which entails claims that are in conflict with $\mathbf{e}$. Technically, refuting samples should go beyond basic negations such as "Rome is not in Italy." They should instead be adept at capturing nuanced factual contradictions, e.g., "Rome is in France", "There are two Coliseums in Rome." Obtaining such variety in claims remains an open research question.

Challenge II: Low-Budget Environment. In low-resource settings, restrictions such as commodity hardware infrastructure can affect model supervision and performance (Parida and Motlícek, 2019; Moro and Ragazzi, 2022, 2023; Huh and Ko, 2023; Moro et al., 2023a,b,c). In the era of LLMs, the investigation of flexible and scalable solutions is being neglected despite their high social impact (Tamkin et al., 2021). Developing FC systems capable of scaling and adapting to diverse user needs and scenarios is imperative.

4 Method

We introduce UNOWN (Figure 3), a novel framework to automate the production of FC training data. In a first step, $\mathbf{e}$ is created from the input $\mathbf{d}$ (*evidence selection*). Then, $\mathbf{e}$ is used to generate supporting or refuting claims (*claim generation*).

4.1 Evidence Selection

Anchor Creation. The evidence construction process begins by creating a textual anchor $\mathbf{a}$. We distinguish two settings. **Text-only:** $\mathbf{a}$ is a randomly selected sentence from the document $\mathbf{d}$. **Text+Table:** we combine textual and tabular data to determine $\mathbf{a}$. In alignment with the text-centric vision of previous works (Berrios et al., 2023; Tan et al., 2023; Zeng et al., 2023), we fine-tune T5-large (780M parameters) (Raffel et al., 2020) on TOTTO (Parikh et al., 2020), a table-to-text dataset. We sample table cells following a distribution based on the observed tabular evidence size in the FEVEROUS training set (i.e., $[2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8]$) to generate $\mathbf{a}$ (text) by inference, unifying the data modalities.[6] The prompt uses cell values and includes contextual details such as table headers and the document title to maintain coherence (see Figure 4). This approach eases claim generation but still leaves the question of how to select evidence.

Evidence Completion. Once $\mathbf{a}$ is created, we propose two alternative strategies to complete the evidence. **Random:** we pick k random sentences from $\mathbf{d}$. Various topics may exist within $\mathbf{e}$, as the information chosen may not be aligned. **Semantic Consistency:** $\mathbf{e_s}$ is built by concatenating the k sentences from $\mathbf{d}$ that semantically align the most with $\mathbf{a}$, preserving the topic coherence. As in Liu et al. (2023), we use cosine similarity after T5 encoding.

We expand on important clarifications.

1. In text-only scenarios, $\mathbf{e_t}{=}\emptyset$ and $\mathbf{e}$ consists of a set of sentences. In text+table scenarios, $\mathbf{e}$

[5] The label *Not Enough Information* is excluded due to its rarity, accounting for only 3% of instances in FEVEROUS.

[6] Cell extraction is a consolidated practice for evidence retrieval in table-based factuality predictors (Aly et al., 2021).

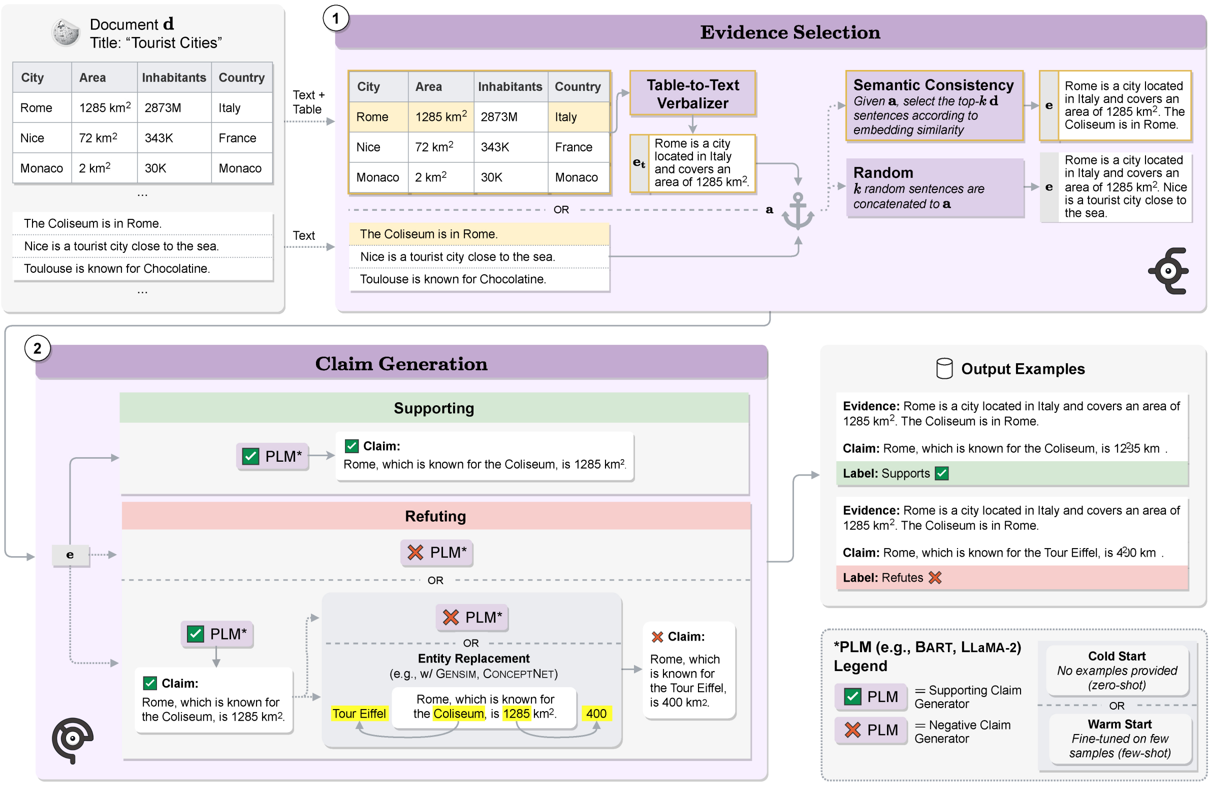

Figure 3: UNOWN pipeline. The input document $\mathbf{d}$ consists of sentences and optional tables. (1) When both modalities are used, we obtain $\mathbf{e_t}$ with a cell sampling and verbalization process. From $\mathbf{e_t}$, different strategies can be used to determine $\mathbf{e_s}$ and complete $\mathbf{e}$; in a text-only approach ($\mathbf{e_t} = \emptyset$), $\mathbf{e}$ is established after sentence sampling. (2) We generate supporting and refuting claims using PLMs. Non-continuous lines and arrows delineate alternatives.

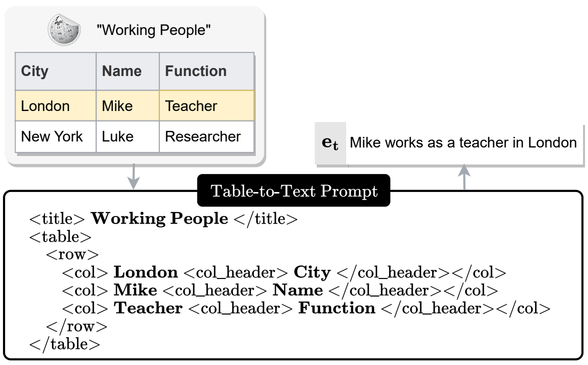

Figure 4: Verbalization of a subset of tabular cells.

comprises sentences and a verbalized representation of tabular cells. We overwrite $\mathbf{e}$ by prefixing the $\mathbf{d}$ title for context with special `<title>` and `<evidence>` token delimiters. Concatenation enables cross-attention among the page title, cells, and sentences.

2. k is drawn randomly from a distribution of $[1, 1, 2, 2, 2, 3, 3, 4, 5]$, selected based on patterns observed in the FEVEROUS training set.
3. We emphasize that constructing $\mathbf{e}$ from $\mathbf{e_t}$ to $\mathbf{e_s}$ using a single verbalization step is the most practical approach, avoiding the complexities of reverse operations.

4.2 Claim Generation

Fine-tuning models on data aligned with the target task has proven effective in enhancing performance (Gururangan et al., 2020). Practically, users can expect access to *external* data from related FC applications and a limited number (e.g., 10, 100) of *internal* human samples specific to the downstream task. Given this context, we define the following concepts to guide our methodology.

Warm-start: *external* examples are available for preliminary training (i.e., $\mathbf{e} \rightarrow c$).

Cold-start: no *external* data is available.

Few-shot learning: *internal* examples are accessible for specialized fine-tuning (regardless of warm/cold start).

Refuting Claims. Generating refuting claims comes with additional intricacies. We recognize two main paths to avoid introducing a strong lexical bias in the artificial training samples, such as basic negation types. **Direct Refusal:** we use a PLM that can directly transform e into a refuting claim, ensuring a direct and straightforward method. **Two-Step Approach:** we summarize e into a supporting claim and apply a targeted modification to flip its meaning. This involves either using *Direct Refusal* with the supporting claim or employing ER, where keywords are strategically swapped with antonyms or related terms from a KB.

5 Experimental Setup

Our focus is on evaluating the veracity component of the FC process during test time, where models are provided with gold evidence alongside the claim for verification. To achieve this, we address the following research questions:

- Q1 Are our generated artificial examples effective for training FC models?
- Q2 Which evidence selection strategy yields the best performance?
- Q3 What method is recommended for generating refuting claims?
- Q4 To what extent does the efficacy of synthetic examples generalize across various domains?
- Q5 How many internal dataset-specific samples are necessary for few-shot learning to bootstrap the downstream FC model successfully?

Datasets. In warm-start scenarios, we use a subset of 10K positive and 10K negative human examples from FEVER (Thorne et al., 2018), a collection of claim–evidence pairs based on Wikipedia. As the leading FC benchmark, we take **FEVEROUS** (Aly et al., 2021), an extension of FEVER with more complex claims enriched with tabular evidence (with no overlap between the two corpora). Since the original test set is private and lacks gold labels and evidence for the claims, we used the provided validation set as our test set for evaluation. We then divided this set into two subsets: one containing claims based solely on textual evidence, and another containing claims that require both textual and tabular evidence (we exclude claims relying only on tables). To assess generality, we include **SCIFACT** (Wadden et al., 2020), a dataset of expert-written claims paired with evidence from scientific papers abstracts. For the same rationale applied to FEVEROUS, we used the original validation set as our test set. Finally, we release **MMFC**, a new multi-modal FC corpus. Mechanically, we sample 2000 instances from MULTIMODALQA (Talmor et al., 2021), a QA

Dataset	Use Case	Veracity Labels†	Claim Length	Evidence Sent./Cells*
Claim Generation				
FEVER 🗎	Warm-Start	10K ✅ / 10K ❌	8.1	2.4/0
FEVEROUS 🗎 ‡	Few-Shot Learning	0.1K ✅ / 0.1K ❌	27.7	2.1/0
Fact Verification				
FEVEROUS 🗎	Train	16.2K ✅ / 12.7K ❌	27.7	2.2/0
	Test	1.5K ✅ / 1.7K ❌	27.1	2.1/0
FEVEROUS 🗎 + ⊞	Train	15.8K ✅ / 2.3K ❌	26.3	1.6/5.4
	Test	1.5K ✅ / 0.5K ❌	25.2	1.6/4.4
SCIFACT 🗎	Train	0.3K ✅ / 0.2K ❌	12.1	2.1/0
	Test	0.2K ✅ / 0.1K ❌	12.3	1.8/0
MMFC 🗎 + ⊞	Test	0.25K ✅ / 0.25K ❌	21.3	1.5/1.9

† ✅ = supporting claims; ❌ = refuting claims. * Average.
‡ 0.01K and 10K variants are also explored.

Table 2: Dataset statistics. Top area: data eventually employed to align a PLM to the claim generation task before using it. Bottom area: evidence-claim-verdict triplets used to train the fact verification model (UNOWN-generated data) or test it (gold data).

dataset requiring joint reasoning over text, table, and images. In our sampling procedure, we filter out instances requiring visual grounding. Then, we perform few-shot in-context learning with GPT-4-TURBO to transform each question–answer pair into a claim paired with text+table evidence. Finally, we carefully review all examples through human verification to ensure that all reference claims were qualitatively accurate and correctly labeled. Dataset statistics are provided in Table 2. See the Appendix for details.

Metrics. We assess FC predictions using Accuracy and F1 scores ($[0, 1]$; higher is better), distinguishing between *Supports* and *Refutes* labels. We validate models on the test sets after training with artificial and human examples. We finally evaluate the logical relationship between each evidence–claim pair with a DEBERTA cross-encoder (Reimers and Gurevych, 2019) pretrained on natural language inference (NLI) tasks to classify pairs as *Entailment*, *Contradiction*, or *Neutral*.

Claim Generation Models. As SLM, we use models built on BART (Lewis et al., 2020). For supporting claims, we employ the large version (400M parameters). For refuting claims, we utilize two variants: BART-large and BARTNEG (Lee et al., 2021), a specialized BART-base model (140M parameters) trained on parallel and opposing claims from the WIKIFACTCHECK dataset (Sathe et al., 2020).[7] As LLM, we operate with LLAMA-2-7B (Touvron et al., 2023), opting for QLoRA (Dettmers et al.,

[7] Although BARTNEG has already undergone a warm-start process, applying warm start with FEVER is still necessary to deal with multi-sentence input and language style adaptation.

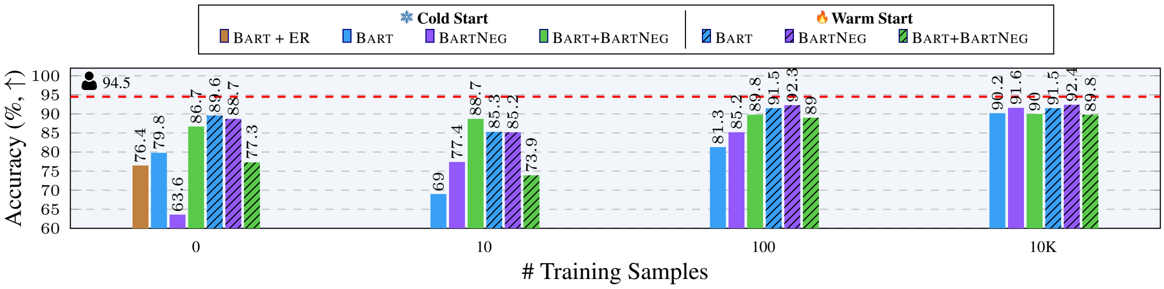

Figure 5: Accuracy scores on FEVEROUS by varying the number of its training samples. Dashed bars indicate the use of external fine-tuning on FEVER. The red dashed line represents the accuracy obtained by human data.

2023) adapter fine-tuning (the prompt template is provided in the Appendix). We stress that refuting claim generation can be obtained by: (i) running these models directly on **e**; (ii) applying these models to the claim returned by a supporting model. Training is done independently for the two claim types; details are reported in the Appendix.

Entity Replacement Methods. As a baseline method, and to show the generality of our framework, we adapt the pipeline proposed by Wright et al. (2022) to domain-agnostic resources, studying three alternative refuting claim generation procedures. **(1)** We prompt FLAN-T5-large (780M parameters) (Wei et al., 2022) with *"Answer the following question. Can you give me an antonym of* $\{\{\mathbf{w}\}\}$*?"*, where $\mathbf{w}$ is a word of length ≥ 4 characters randomly chosen for replacement. **(2)** We use the GENSIM library (Rehurek and Sojka, 2011) to calculate a similarity matrix between the words in the supporting claim. The matrix is subsequently used to build a frequency ranking, aiding in deciding which word to replace (least common, most common, random). Denoting the chosen item as $\mathbf{w}$, words having similarity > 0.7 to $\mathbf{w}$ are substituted with a similar but distinct word as per WORDNET (Miller, 1995). **(3)** We use CONCEPTNET (Speer et al., 2017) to identify a set of concepts closely related to each word in the claim. We build a claim-level frequency ranking on the intersection of word-level concepts. Then, we replace $\mathbf{w}$ according to `antonym` relationships.

Fact-Checking Models. We assess the impact of our synthetic training examples on accurately predicting the verdict label of an input claim given a set of evidentiary sentences. To achieve this, we choose optimal classification models for the benchmarks at hand, keeping their weights and hyperparameters unchanged. For FEVEROUS and MMFC, we use ROBERTA (Liu et al., 2019) with a linear layer on top. For SCIFACT, we employ MULTIVERS (Wadden et al., 2022) with a shared encoding of the claim and input context.

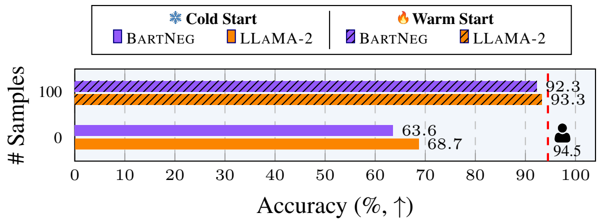

Figure 6: Accuracy scores on FEVEROUS with training examples generated by LLAMA-2.

6 Results and Discussion

6.1 Q1 Quality of Generated Claims

SLMs. We evaluate how UNOWN training examples generated by small models contribute to a downstream FC system by measuring performance on the FEVEROUS test set (Figure 5). In the worst-case scenario (cold-start, zero-shot learning), the highest accuracy achievable by UNOWN is 86.7 with BART-large used for the generation of both supporting and refuting claims. When leveraging human training instances, the results show a consistent boost in performance. In fact, accuracy climbs to 92.3 with warm start and just 100 internal target examples, using BART-large for supporting claims and BARTNEG for direct refusal—close to the accuracy achievable with human-annotated data (94.5).

LLMs. Figure 6 looks at how the claims generated by LLAMA-2 stack up against those inferred by the best SLM setup. The accuracy propelled by LLAMA-2 claims, after training on 100 internal examples, is 93.3, outperforming the small solution by a single point. Therefore, incorporating LLMs does not appear essential in the UNOWN pipeline, favoring BART-based models for their

	Method	Entail.†	Contrad.‡	Neutral
	✅ Supporting			
👤	HUMAN	75.00	4.00	21.00
🤖	BART	74.57	4.90	20.53
	LLAMA-2	71.92	3.95	24.12
	❌ Refuting			
👤	HUMAN	3.00	77.00	30.00
🤖	BART	10.43	56.00	33.57
	LLAMA-2	11.23	37.82	50.95
	ENTITY REPLACEMENT	36.05	40.70	23.26

† [0, 100]. ✅ : ↑ (higher is better). ❌ : ↓ (lower is better).
‡ [0, 100]. ✅ : ↓ (lower is better). ❌ : ↑ (higher is better).

Table 3: The quality of the generated claims in FEVEROUS based on NLI scores (text-only scenario).

Challenge	👤	🤖
COMBINING TABLES AND TEXTS	**93.0**	**93.0**
SEARCH TERMS NOT IN CLAIM	94.0	**96.0**
MULTI-HOP REASONING	**96.0**	92.0
NUMERICAL REASONING	**93.0**	88.0
ENTITY DISAMBIGUATION	**88.0**	79.0
OTHER	**96.0**	93.0

Table 4: Accuracy of the FC model on challenge-specific subsets of FEVEROUS when trained on human-annotated or UNOWN data. The best results are in bold.

superior effectiveness–efficiency trade-off (see Table 8 in the Appendix for efficiency statistics).

NLI. To gain additional insight into the generated claims, we compute the NLI prediction score between claims and evidence. Table 3 shows that, for supporting claims, UNOWN's examples closely resemble the score distribution in their human-written counterparts. Yet, in the refuting examples generated by UNOWN, the percentage of entailed claims surpasses that of human-generated ones, highlighting the greater difficulty in creating refuting examples compared to supporting ones. We observe that the ER baseline performs the worst.

Challenges of Claim Verification. We evaluate the effectiveness of our data generation method across challenge categories defined by Aly et al. (2021). Specifically, we compare the performance of an FC model trained on UNOWN data versus human-crafted data on different subsets of the FEVEROUS test set, each focused on a particular challenge. As shown in Table 4, the FC model trained on our data performs competitively in several categories, such as "Combining Tables and Texts" and "Search Terms Not in Claim," even outperforming the model trained on human-generated data. While the FEVEROUS-trained model holds a slight advantage in areas like "Multi-hop Reasoning," "Numerical Reasoning," and "Entity Disambiguation," our approach radically reduces the need for expensive and time-consuming human annotation.

Human Evaluation. We perform a qualitative analysis to investigate the quality of the claims generated by UNOWN. We randomly sample 50 instances from the FEVEROUS training data (25 supporting, 25 refuting). Taking into account the expense associated with careful human evaluation and the central role of text as our unified modality, we accord priority to text-only evidence. Each instance is presented with its original human-selected evidence and the corresponding claim. To maintain fairness, we condition our models on this evidence and generate synthetic claims using our best-performing models: the warm-started BART-large for supporting claims and BARTNEG for refuting claims. After manually verifying the correctness of the assigned label, which were accurate for all 50 claims, we enlist the expertise of three external annotators with strong NLP and FC backgrounds to evaluate the claims. In a blind review process, we provide them with the evidence and the two claims (original and generated) in randomized order. Following a direct comparison assessment, which has proven to be more reliable and sensitive than rating scales (Kiritchenko and Mohammad, 2017) and has been used to evaluate abstractive summaries (Fabbri et al., 2019; Moro et al., 2023d; Ragazzi et al., 2024) and answers (Moro et al., 2024), we ask the annotators to determine which claim is the best with respect to two dimensions: *clarity* (effective communication of the intended meaning with a good sentence structure, fluency, and English precision) and *coherence* (semantic connection to the evidence). They are also given the option to declare a tie if they perceive the quality of the claims to be comparable. To aggregate the annotations, we employ a majority voting approach and calculate Cohen's κ coefficient to gauge the agreement between annotators and the majority voting label. The coefficient value of 0.613 indicates a substantial level of agreement, enhancing the reliability of our analysis. As illustrated in Figure 7, the results reveal an interesting landscape. Out of the 50 paired claims, annotators found 35 to be of comparable quality. In 10 cases, the original FEVEROUS claims were deemed superior, while in 5 cases, the claims generated by UNOWN were judged to be of higher quality.

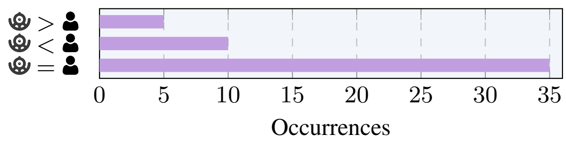

Figure 7: Human evaluation results on 50 claims.

Method	Text			Text+Table		
	Acc.	F_1 ✅	F_1 ❌	Acc.	F_1 ✅	F_1 ❌
FEVEROUS						
EVID. + CLAIM	94.50	94.92	95.30	82.09	87.83	66.12
EVID. + CLAIM	92.40	92.10	92.70	84.59	90.53	58.69
RAND. w/ GOLD + C	92.30	91.70	92.70	81.81	87.96	62.89
RAND. w/o GOLD + C	85.11	86.07	84.01	84.43	89.90	66.03
SEM. CONSIST. + C	88.33	88.21	88.44	86.83	91.73	67.65
MMFC						
EVID. + CLAIM	-	-	-	87.60	88.17	86.97
EVID. + CLAIM	-	-	-	76.00	75.61	76.38

Table 5: Evidence selection comparison in FEVEROUS and MMFC. Methods use BART and BARTNEG to create supporting and refuting claims, respectively. Models are trained on FEVER and then on 100 dataset samples.

Overall, the generated examples (see the Appendix) prove to be sufficiently effective for training FC models, yielding quantitative results in a 2-point margin in absolute accuracy compared to those achieved by a crowd of annotators.

6.2 Q2 Evidence Selection

We study the impact of alternative evidence selection methods. We report two experiments using FEVEROUS training examples: one with text-only evidence and another with text+table evidence; test datasets are filtered according to the scenario. For every human example, referred to as "gold," we execute our best BART model with four alternative evidence selection strategies. **Human evidence**, where we use the original evidence handpicked by the annotators. **Random with gold**, where the number of selected sentences matches the human example, but the actual cells and sentences are chosen randomly from d. **Random without gold**, where the number of retrieved sentences k, after anchor definition, is drawn from the distribution presented in Section 4.1. **Semantic consistency**, where textual evidence is retrieved using embedding similarity to the table verbalization (see Figure 4).

Table 5 shows accuracy and F1 results. The influence of evidence is evident. The use of human evidence allows UNOWN to produce examples that match nearly the human upper bound. In the text+table setting, we achieve even higher scores for supporting claims, confirming the quality of our claim generator. In the text-only scenario, performance is optimal when guided by the cardinality of human gold evidence, with random selection surpassing semantic consistency. In text+table, semantic consistency outperforms both random selection and original human examples in all metrics. We observe that human annotators struggle to annotate tabular data accurately, making mistakes that mislead the classifier. This is also reflected in the generally lower results for text+table compared to the text-only scenario.

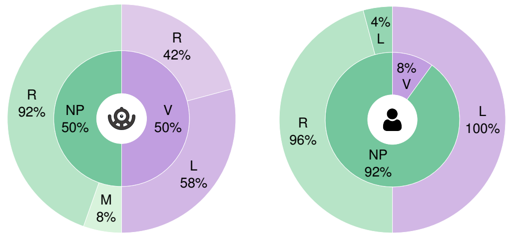

Figure 8: Human annotation on negation artifacts.

Table 5 also shows the results for MMFC. In this dataset, all claims involve text and tabular data and we only have human gold evidence for the original claim. We explain the lower quality results for UNOWN because the warm start includes examples from FEVER, which are different from those in MMFC (see the Appendix for examples), possibly introducing a negative bias.

6.3 Q3 Refuting Claims

We show how the FC performance varies with different types of *Refutes* generated claims in a quantitative analysis and then in a qualitative user study.

Quantitative. We identify FLANT5 as the best ER method (the results are shown in Figure 12 in the Appendix); unless otherwise specified, we use ER to denote this baseline approach. Figure 5 includes the impact of various negation strategies on the accuracy of the target task. In cold start, the combination of BART and BARTNEG using the two-step approach is effective, while the results are subpar with ER, which fails to make refuting claims, possibly due to limitations in content replacement without adequate rewording. As anticipated, starting with a warm start is beneficial, resulting in the highest accuracy with 0 and 100 training samples.

Qualitative. We perform a human analysis to evaluate the negation techniques used to refute claims. We adhere to the negation taxonomy outlined in previous studies (Zafra et al., 2020; Dobreva and Keller, 2021). Rigorously, we use two main nega-

	Method	Accuracy	F_1 ✅	F_1 ❌
👤	HUMAN TRAINING DATA	84.62	81.05	85.71
🤖❄️	ENTITY REPLACEMENT	55.33	57.27	16.51
	BARTNEG	65.08	53.79	65.96
🤖🔥	ENTITY REPLACEMENT	54.00	57.63	1.98
	BARTNEG	74.50	73.53	73.45

Table 6: Strategies for refuting claim generation on SCIFACT; models use BART to create supporting claims. In warm scenarios, models are fine-tuned on FEVER.

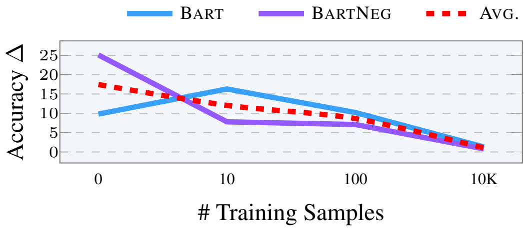

Figure 9: The FEVEROUS's average Δ accuracy improvement when shifting from cold to warm.

tion types, namely *Verbal Negation* (V) and *Noun Phrase Negation* (NP). Each is classifiable in three subclasses, including *Lexical* (L), where the negation is expressed with new words or phrases that alter the sentence meaning (e.g., 10 papers→**more than** 10 papers), *Morphological* (M), where the form of the word is modified through morphemes (e.g., legal→**il**legal), and *Replacement* (R), where a phrase is swapped for another with a different meaning (e.g., 1995→1997). Given these classes, three annotators (selected among the authors) evaluated 30 refuting claims from the original FEVEROUS training dataset and 30 refuting claims generated by UNOWN. The final category is identified by majority voting over the three suggested labels; the Cohen's κ coefficient is 0.91, which shows very high agreement among annotators. The results of the study are illustrated in Figure 8, allowing a comparison of annotation distributions between the two sets of examples (UNOWN vs. human). UNOWN produces an even distribution of refuting claims, encompassing both noun phrases and verbal structures, whereas humans tend to prefer noun phrases. Both UNOWN and humans favor the replacement strategy for noun phrases and the lexical strategy for verbs. In both scenarios, the ranking of classes and subclasses remains consistent, indicating that UNOWN produces a range of negation types comparable to those observed in a human-crafted corpus.

6.4 Q4 Checking Scientific Claims

We measure the quality of the FC system trained with UNOWN examples in a different domain. Due to the lack of heterogeneous datasets such as FEVEROUS, we use the text-only scientific corpus SCIFACT. Table 6 confirms the analysis outcome on FEVEROUS. Human data achieve the best results, followed by UNOWN with the warm-started BART. We explain the greater result gap between humans and UNOWN because the warm start includes only examples from FEVER. Again, BARTNEG leads to better results with respect to ER. We posit that low F1 refuting scores (i.e., 1.98, 16.51) stem from FLAN-T5's pre-knowledge bias, which may not adequately align with scientific subjects.

6.5 Q5 Bootstrapping: Cold vs. Warm Start

We measure the impact of the examples used to fine-tune the models. As shown in Figure 5, Figure 6, and Table 6, a warm-start approach improves the quality of the generated data. More precisely, Figure 9 shows the average Δ accuracy improvement when shifting from cold to warm in FEVEROUS. We observe a decrease in Δ as the number of internal samples from the target dataset increases, highlighting the beneficial contribution of using external related data as a guide source of knowledge. Also SCIFACT exhibits an increase in accuracy for BARTNEG in the warm approach.

7 Conclusion

We introduced UNOWN, a domain-agnostic framework to automatically generate training examples for fact-checking systems, bypassing the costly task of manually annotating large volumes of data. UNOWN fits both structured and unstructured data to compile textual claims that support or refute the evidence provided. It also accommodates several solutions for evidence selection and claim generation to adapt to different scenarios. We evaluated our framework using three datasets that deal with general-domain and scientific contexts. The results indicate that our synthetic examples exhibit a quality comparable to that of expert-labeled data, showing the practicality and efficacy of our framework. Quantitative and human evaluation also register that our refuting examples have high variety, comparable to human-generated ones.

Limitations

Although UNOWN is a promising step forward, some research directions remain unexplored. First, our generation process lacks coverage of certain examples within the long tail, e.g., mathematical operations, such as the premise "Paul is 2 years younger than Mary." We consider using a solution in which more intricate patterns are generated as queries over relational tables (Bussotti et al., 2023). Second, once models have been trained with instances from UNOWN, we could set up active learning algorithms to guide our methods in generating examples that effectively enhance performance on the test set (Zhang et al., 2022). Third, while the considered datasets contain well-crafted claims, real-world claims can often be incomplete—lacking context and presenting ambiguity in relation to the evidence (Glockner et al., 2024)—or require multi-modal evidence that extends beyond text and tables (Akhtar et al., 2023). Furthermore, reasoning over multiple pieces of evidence from diverse sources may also be necessary. Finally, the selection of a specific model for generating supporting or refuting claims can result in diverse fact-checking challenges that may vary in their alignment with the target dataset. For instance, the FEVEROUS test set contains instances that demand robust multi-hop reasoning abilities, whereas other benchmarks might require advanced numerical reasoning skills. This observation helps explain why, despite using identical models for synthetic data generation, the text+table performance achieved by ROBERTA on MMFC after training on synthetic data is less promising compared to its performance on FEVEROUS. These findings underscore the importance of future research efforts to explore methods for better aligning synthetic data with the characteristics of specific target datasets. Future endeavors could also consider the evidence retrieval stage (Frisoni et al., 2022), cross-domain FC (Kao and Yen, 2024; Domeniconi et al., 2014), and knowledge extracted from unlabeled corpora (Frisoni and Moro, 2020) to force the generation of cross-document claims.

Ethics and Impact Statement

Although fact-checking systems like UNOWN enhance information integrity and combat misinformation, it is essential to ensure their responsible and beneficial use in society. UNOWN aims to efficiently generate training instances, yet it is needed to rigorously validate and supervise the synthetic examples to ensure that they accurately represent real-world scenarios without introducing inadvertent biases. Moreover, high-resource language models demonstrate limited effectiveness when applied to low-resource language data (Huang et al., 2023). Similarly to various domains within NLP that depend on meticulously constructed datasets, fact-checking contributions have mainly focused on a few high-resource languages, such as English and Chinese (Zarharan et al., 2021). As this could skew perceptions of automated fact-checking advancements, future studies should prioritize advances in false claim detection for low-resource languages.

Acknowledgements

This research is partially supported by (i) the ANR project ATTENTION (ANR-21-CE23-0037), (ii) the AI-PACT project (CUP B47H22004450008 and B47H22004460001), (iii) the Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE—DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (iv) the PNRR—M4C2—Investment 1.3, Extended Partnership PE00000013, FAIR—Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program, (v) the European Commission and the Italian MIMIT through the Chips JU TRISTAN project (G.A. 101095947).

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5430–5448. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the LENS of natural language. *CoRR*, abs/2306.16410.

Jean-Flavien Bussotti, Enzo Veltri, Donatello Santoro, and Paolo Papotti. 2023. Generation of training examples for tabular natural language inference. *Proc. ACM Manag. Data*, 1(4):243:1–243:27.

Kevin Matthe Caramancion. 2023. News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *CoRR*, abs/2306.17176.

John M. Carey, Andrew Markus Guess, Peter John Loewen, Eric Merkley, Brendan Nyhan, Joseph B. Phillips, and Jason Reifler. 2022. The ephemeral effects of fact-checks on covid-19 misperceptions in the united states, great britain and canada. *Nature Human Behaviour*, 6:236 – 243.

Vincenzo Carrieri, Sophie Guthmuller, and Ansgar Wübker. 2023. Trust and covid-19 vaccine hesitancy. *Scientific Reports*, 13.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

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. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314.

Radina Dobreva and Frank Keller. 2021. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, pages 350–362. Association for Computational Linguistics.

Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. 2014. Iterative refining of category profiles for nearest centroid cross-domain text classification. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers*, volume 553 of *Communications in Computer and Information Science*, pages 50–67. Springer.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Giacomo Frisoni and Gianluca Moro. 2020. Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In *Data Management Technologies and Applications - 9th International Conference, DATA 2020, Virtual Event, July 7-9, 2020, Revised Selected Papers*, volume 1446 of *Communications in Computer and Information Science*, pages 293–318. Springer.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12365–12394. Association for Computational Linguistics.

Taehun Huh and Youngjoong Ko. 2023. Efficient framework for low-resource abstractive summarization by meta-transfer learning and pointer-generator networks. *Expert Syst. Appl.*, 234:121029.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: Multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10891–10902, Torino, Italia. ELRA and ICCL.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 465–470. Association for Computational Linguistics.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheon-Eum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3181–3185. ACM.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *CoRR*, abs/2305.14623.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. *CoRR*, abs/2305.03695.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6826–6833. AAAI Press.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11085–11093. AAAI Press.

Gianluca Moro and Luca Ragazzi. 2023. Align-then-abstract representation learning for low-resource summarization. *Neurocomputing*, 548:126356.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023a. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14417–14425. AAAI Press.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023b. Graph-based abstractive summarization of extracted essential knowledge for low-resource scenarios. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1747–1754. IOS Press.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Giacomo Frisoni, Claudio Sartori, and Gustavo Marfia. 2023c. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*, 23(7):3542.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Lorenzo Molfetta. 2023d. Retrieve-and-rank end-to-end summarization of biomedical studies. In *Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, October 9-11, 2023, Proceedings*, volume 14289 of *Lecture Notes in Computer Science*, pages 64–78. Springer.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Fabian Vincenzi, and Davide Freddi. 2024. Revelio: Interpretable long-form question answering. In *The Second Tiny Papers Track at ICLR 2024, Tiny*

Papers @ ICLR 2024, Vienna, Austria, May 11, 2024. OpenReview.net.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 476–483. Association for Computational Linguistics.

Shantipriya Parida and Petr Motlícek. 2019. Abstract text summarization: A low resource challenge. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5993–5997. Association for Computational Linguistics.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1173–1186. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Luca Ragazzi, Gianluca Moro, Stefano Guidi, and Giacomo Frisoni. 2024. Lawsuit: a large expert-written summarization dataset of italian constitutional court verdicts. *Artificial Intelligence and Law*.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2116–2129. Association for Computational Linguistics.

Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 1736–1746. ACM.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6874–6882. European Language Resources Association.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, Nick Ryder, Alex Paino, Qiming Yuan, Clemens Winter, Ben Wang, Mo Bavarian, Igor Babuschkin, Szymon Sidor, Ingmar Kanitscheider, Mikhail Pavlov, Matthias Plappert, Nik Tezak, Heewoo Jun, William Zhuk, Vitchyr Pong, Lukasz Kaiser, Jerry Tworek, Andrew Carr, Lilian Weng, Sandhini Agarwal, Karl Cobbe, Vineet Kosaraju, Alethea Power, Stanislas Polu, Jesse Han, Raul Puri, Shawn Jain, Benjamin Chess, Christian Gibson, Oleg Boiko, Emy Parparita, Amin Tootoonchian, Kyle Kosic, and Christopher Hesse. 2022. Introducing chatgpt.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *CoRR*, abs/2102.02503.

Wang-Chiew Tan, Yuliang Li, Pedro Rodriguez, Richard James, Xi Victoria Lin, Alon Y. Halevy, and Wen-tau Yih. 2023. Reimagining retrieval augmented language models for answering queries. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6131–6146. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Ozan Tonguz, Yiwei Qin, Yimeng Gu, and Hyun Hannah Moon. 2021. Automating claim construction in patent applications: The cmumine dataset. In *Proceedings of the Natural Legal Language Processing Workshop 2021, NLLP@EMNLP 2021, Punta Cana, Dominican Republic, November 10, 2021*, pages 205–209. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Enzo Veltri, Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Data ambiguity profiling for the generation of training examples. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 450–463. IEEE.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 61–76. Association for Computational Linguistics.

Nancy Xin Ru 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@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 317–326. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2448–2460. Association for Computational Linguistics.

Shuo Yin, Weihao You, Zhilong Ji, Guoqiang Zhong, and Jinfeng Bai. 2024. Mumath-code: Combining tool-use large language models with multi-perspective data augmentation for mathematical reasoning.

Salud María Jiménez Zafra, Roser Morante, María Teresa Martín-Valdivia, and Luis Alfonso Ureña López. 2020. Corpora annotated with negation: An overview. *Comput. Linguistics*, 46(1):1–52.

Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. 2021. Parsfever: a dataset for farsi fact extraction and verification. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 99–104. Association for Computational Linguistics.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhisong Zhang, Emma Strubell, and Eduard H. Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6166–6190. Association for Computational Linguistics.

Appendix

Multi-Modal Evidence. We conduct an ablation study aimed at evaluating the importance of each evidence modality for table+text FC instances (Table 7). When text or cells are excluded from the evidence in the test data, accuracy and F1 scores for the FC model drop significantly.

Test set	**Accuracy**	**F_1 ✅**	**F_1 ❌**
STANDARD TEST DATA	86.9	91.7	67.6
ABLATED TABLES	57.6	66.0	43.9
ABLATED SENTENCES	62.8	71.5	46.3

Table 7: Results on three different test sets: the gold test set, the same test set with ablated tables in evidence, and the same test set with ablated sentences in evidence. Training data is always based on the warm start and the BART/BARTNEG combination.

Environment. We run each experiment on a cluster of OS Linux workstations with a single Nvidia GeForce RTX3090 Turbo GPU of 24 GB VRAM. UNOWN is developed using PyTorch (Paszke et al., 2019) and the HuggingFace library (Wolf et al., 2019) (seed set to 42 for reproducibility).

Experimental Setting. To train BART, we set the following hyperparameters: learning_rate=$1e^{-4}$, batch_size=16, and epochs=20; for LLAMA-2, we use 4-bit nested quantization, r=8, α=32, batch_size=1, and epochs=3. For inference, we adopt beam search (num_beams=5) and nucleus sampling (top_p=0.01, top_k=40, temp=0.15) for BART and LLAMA-2, respectively.

Execution Times. Table 8 reports the train and inference time per claim for the claim generation task. The benefit of smaller models is evident during inference. We also report the average time required to generate an example in terms of evidence selection. The total time of about 6 seconds per claim is in contrast to the time and effort required by a human to craft a comparable example.

Examples. Tables 9, 10, and 11 report examples of textual claims generated by our system with different models given the same original evidence. The human-written claim is provided for comparison. We note that many claims generated by BARTNEG with *Refutes* labels do not contain the word "never". To illustrate:

- *"In the 2006-07 San Jose Sharks season, the team scored 107 goals, 183 assists, and 1*

Model	Task	sec/Claim
Claim Generation		
BART	`Train/Infer.`	1.92 / 0.12
BARTNEG	`Train/Infer.`	1.01 / 0.08
LLAMA-2	`Train/Infer.`	1.98 / 2.10
Table-to-Text		
T5-ToTTo	`Infer.`	0.75
Evidence Selection (Semantic Consistency)		
T5	`Tokeniz. + Distance`	5.43

Table 8: Time consumption for different tasks.

Shutout." Here, the real numbers are 107, 283 and 5.

- *"Karyn Kupcinet, who died on June 2, 1963, appeared on The Donna Reed Show and The Gertrude Berg Show, 1999."* Here, the actual day is November 28, 1963.
- *"Rihanna had a live performance at the Super Bowl in 2012."* Here, the actual singer is Madonna.

These examples showcase the variability of our generated claims, ensuring that the models trained on our data must learn robust patterns beyond simple negations and manage hard negative cases from a semantic viewpoint. Additionally, we acknowledge the presence of several generated claims with *Supports* labels that contain the word "never", further requiring the ability to capture diverse linguistic patterns. For instance *"Bruce Johnston's song 'I Write the Songs' never charted."*

Claim Generation Prompts. Prompt tuning experiments proved the marginal role of few-shot in-context learning. We then opted for a simpler and reproducible zero-shot approach, also fairer to small models, as reported in Figure 10.

Claim Generation Prompt

Write a claim that uses the following evidence.
Write a negative claim, i.e., false with regard to the following evidence.

Evidence:
<title> {{**d** title}} <evidence> {{**e**}}

Claim:
y

Figure 10: Instruction tuning prompt template for claim generation. The highlighted part is used for loss computation. Green and red colors denote alternative instructions for supporting and refuting targets, respectively.

Numerical Reasoning. The FEVEROUS datasets contains several reasoning examples. Examining its test set for table+text, we reviewed 100 claims and identified only 7 instances requiring reasoning through cell aggregation. Consequently, we investigated how our system was able to deal with them. We compared some examples of human-written claims versus UNOWN generated ones, using the same evidence. We showcase them in Table 12. For each example, we present the claim generated, along the intermediate text it generated from the table. In the first example, we can see that both the table to text system and the final UNOWN simply gave a description of the routes without making any counting. In the second example, even though it appears that our system counted the points, the truth is that this number is present in the original table. In the meantime, the human leveraged this information to create a superlative "scored the most points". In the last examples, again, the T5 Verbalizer simply reports "57%" without trying to convert it to a more subtil information such as "more than half", as the human did. Our system even discard this information and prefers to write a claim about the number of votes.

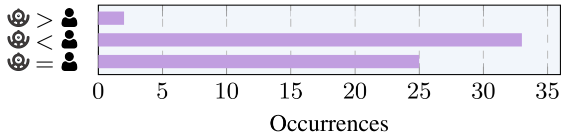

Figure 11: Human evaluation results on 60 claims, involving both tabular and textual evidence.

We emphasize the role played by the T5-large table verbalizer in this observation. The inclusion or exclusion of operations involved in generating textual descriptions associated with the content of sampled cells is largely determined by the training dataset used. The T5 verbalizer is trained on ToTTo (Parikh et al., 2020). An analysis of the distribution of various linguistic phenomena conducted by the dataset's authors reveals that reasoning (including logical, numerical, and temporal) is present in only 21% of the instances. As a second note, even if we provided claims that require operations to verify their accuracy, we cannot expect the final predictor model to handle these operations effectively. As several experiments have demonstrated (Chang et al., 2023), even recent LLMs struggle with basic tasks like averaging. The most recent approach to address this issue is to use external modules, such as Python, to handle the mathematical computations (Yin et al., 2024).

Evidence (d title = Windecker Eagle)	
The Eagle AC-7 Eagle 1 (USAF designation YE-5) is an aircraft that was manufactured by Windecker Industries Data from Jane's All the World's Aircraft 1971.	
✅ **Supporting Claims**	
HUMAN	The Eagle AC-7 Eagle 1 is a military aircraft that was manufactured by Windecker Industries.
LLAMA-2	Windecker Eagle is listed in Jane's All the World's Aircraft 1971-72.
BART	In 1896, Friedrich Baume established a legal practice with A. E. Whitaker, a son of Frederick Whitaker, and won the election.

Table 9: Example from FEVEROUS of UNOWN's supporting claims with textual evidence.

Tabular Evidence (d title = 2010 FedEx Cup Playoffs)	
Bubba Watson [Header : Player]	
Textual Evidence (d title = Bubba Watson)	
Watson was born and raised in Bagdad, Florida, near Pensacola. He played on the golf team at Milton High School, which had featured future PGA Tour members Heath Slocum and Boo Weekley just before he attended. Watson played golf for Faulkner State Community College in nearby Baldwin County, Alabama, where he was a junior college All-American. He transferred to the University of Georgia, the defending NCAA champions, and played for the Bulldogs in 2000 and 2001. As a junior, Watson helped lead the Bulldogs to the SEC title in 2000.	
✅ **Supporting Claims**	
HUMAN	The player who had 1,498 points in the regular season of the 2010 FedEx Cup Playoffs attended Milton High School.
BART	Bubba Watson Watson played in the 2010 FedEx Cup playoffs. Watson was born and raised in Bagdad, Florida, near Pensacola. He played on the golf team at Milton High School, which had featured future PGA Tour members Heath Slocum and Boo Weekley

Table 10: Example from MMFC of UNOWN's supporting claims, with tabular and textual evidence.

Evidence (d title = Stefan Lessard)	
Lessard was born in Anaheim, California, to musician Ron and Jacqueline Lessard. During the 2007 European Tour, Lessard <u>*also began*</u> a video blog on YouTube. In October 2005 he married Jaclyn Burton, and together they have <u>*three daughters*</u>, Hazel Bailey (born March 6, 2004), Flora Jude (born July 13, 2006), and India (born November 2015).	
❌ **Refuting Claims**	
HUMAN	Stefan Kahil Lessard (born on June 4, 1974 **outside** Anaheim, California) completed his junior year of high school, passed the GED examinations in the summer of 1991, and was accepted into the Jazz Studies program at Virginia Commonwealth University.
LLAMA-2	Stefan Lessard has been married to Jaclyn Burton since October 2005. They have **three sons** named Hazel Bailey (born March 6, 2004), Flora Jude (born July 13, 2006) and India (born November 2015).
BART	Stefan Lessard **never began** a video blog on YouTube.

Table 11: Example from FEVEROUS of UNOWN's refuting claims, highlighting negation artifacts (bold) with respect to evidence excerpts (underline).

Example 1	
HUMAN	Lindfield railway station has 3 bus routes, in which the first platform services routes to Emu plains via Central and Richmond and Hornbys via Strathfield.
GENERATED	Lindfield railway station is on the Northern Line (T9), a historical landmark where it has a little bit of accessibility.
TABLE VERBALIZED	Lindfield railway station is served by services to Emu Plains via the Central Railway Station and Richmond via the Northern Railway Station.
Example 2	
HUMAN	The 2006-07 San Jose Sharks season, the 14th season of operation (13th season of play) for the National Hockey League (NHL) franchise, scored the most points in the Pacific Division.
GENERATED	In the 2006-07 San Jose Sharks season, the team scored 183 goals and had a total of 46 Shutouts.
TABLE VERBALIZED	The Anaheim Ducks had 110 points and the San Jose Sharks had 107 points.
Example 3	
HUMAN	During the 2003 Ottawa municipal elections, more than half of the votes in the 8th Zone for the Eastern Ontario Public School Board Trustees seat went to Chantal Lecours.
GENERATED	In the 2003 Ottawa municipal election Denis Chartrand was elected with 760 votes.
TABLE VERBALIZED	Chantal Lecours received 57.84% of the vote in the 2003 Ottawa municipal election."

Table 12: Comparison of FEVEROUS examples and UNOWN's ones on numerical reasoning.

Multi-Modal Human Evaluation. Along the evaluation of Figure 7, we extended our analysis on 100 claims containing a mixture of the Text+Table and Text only settings. We report our analysis in Figure 11. With the frequency of human winning above the frequency of draws, the task here is performed more difficulty by UNOWN. Despite those examples are unpleasant to human, they are efficient in practice for model fine-tuning, as seen in Table 3.

Alternative Entity Replacement Methods. Finally, Figure 12 shows how we identified FLANT5 and random selection as the best combination for the ER method used as our baseline approach.

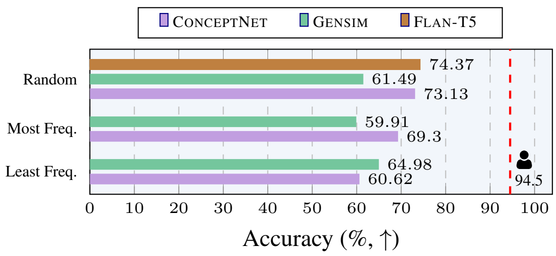

Figure 12: Comparison of different entity replacement methods in FEVEROUS.

MMFC Dataset Building. Supporting claims are generated by prompting GPT-4-TURBO (`gpt-4-turbo-2024-04-09`) as detailed in Figure 13. The examples employed in the few-shot learning process are structured as follows:

- *input* contains the question–answer pair.
- *not optimal output* shows a type of answer to avoid.
- *better output* provides the reference claim.

Refuting claims are generated with the prompt reported in Figure 14. A *why* field clarifies the expected negation behavior and makes explicit the difference between the *not optimal output*, *incorrect output*, and *better output* fields.

We conducted in-depth prompt engineering and manually checked the generated claims, revising them as needed to correct errors.

Question/Answer to Claim Prompt SUPPORTS

Can you make a claim out of this Question/Answer pair? Your answer should only contain the claim. You should add no other information.
Here are some examples of things not to do :
Input : Is the religion with 16.27% of the Canadian Census of 1871 the same religion as the Church of England? No
Not optimal output : The religion constituting 16.27% of the Canadian Census of 1871 is not the Church of England.
Better output : The religion constituting 16.27% of the Canadian Census of 1871 is a religion other than the Church of England.

Input:Which team was Sebastian Svärd on in 2004-05 that played in the 2017 FA Cup final? Arsenal
Not optimal output : Sebastian Svärd was on the Arsenal team in 2004-05.
Better output : Arsenal, the team Sebastian Svärd was on in 2004-05, played in the 2017 FA Cup final

Input:
Question/Answer
Claim:
y

Figure 13: Prompt for the generation of supporting claims from question–answer pairs in MMFC.

Question/Answer to Claim Prompt REFUTES

Can you make a refuted claim out of this Question/Answer pair? Your answer should only contain the claim. The claim should not be based on basic negation

Here are some examples of things not to do and why:
Input : Is mobil 1 the official sponsor for the constructor that had a time/retired of electrical in the Australian Grand Prix race of 2016 f1 team? Yes
Not optimal output : Mobil 1 was not the official sponsor for the constructor that had a time/retired of electrical in the Australian Grand Prix race of the 2016 F1 team
Why : The boolean answer should not cause a poor negation, containing a simple negation
Better output : Google was the official sponsor for the constructor that had a time/retired of electrical in the Australian Grand Prix race of the 2016 F1 team

Input : in the Season victories of 2017 Astana season, where was the grand depart for the 2017 Race when the Location was La Planche des Belles Filles city and country? Düsseldorf, Germany
Not optimal output : The 2017 Race grand depart from Astana season in La Planche des Belles Filles was in Düsseldorf, Germany.
Why : the way the claim is refuted is too subtle
Better output : The 2017 Race grand depart from Astana season was in Paris, France.

Input : When did the home team that had a score of 20-34 in round 7 of the 2018 NRL season enter the NRL? 1988
Incorrect output : The home team that scored 20-34 in round 7 of the 2018 NRL season entered the NRL in 1988.
Why : the generated claim is not false with regards to the question/answer pair
Better output : The home team that scored 20-34 in round 7 of the 2018 NRL season entered the NRL in 1975.

Input : For the religion that has 3,304 females in the Moscow Governorate, what is its primary literary work? Talmud
Incorrect output : The religion with 3,304 female adherents in the Moscow Governorate predominantly follows the Talmud as its primary literary work.
Why : the generated claim is not false with regards to the question/answer pair
Better output : The religion with 3,304 female adherents in the Moscow Governorate predominantly follows the Bible as its primary literary work.

In any case, the text you generate must be false in light of the initial question/answer pair.

Input:
Question/Answer
Claim:
y

Figure 14: Prompt for the generation of refuting claims from question–answer pairs in MMFC.