Do Numbers Speak? Questioning Quantity for Enhanced Fact-Checking

Anonymous ACL submission

Abstract

Despite advancements in automated fact-checking, a notable gap remains in verifying complex claims, particularly those involving numerical data. This underscores the necessity for fact-checking systems that focus on accurately assessing quantitative claims. To address this critical issue, we introduce QLAIM, a pioneering multi-domain dataset focused exclusively on quantitative claims. It includes 33k fact-checked claims featuring various quantitative information, including comparative, statistical, interval, and temporal, accompanied by detailed metadata and supporting evidence. In conjunction with QLAIM, we present Q2FC, a comprehensive fact-checking framework designed to replicate the investigative rigour of human fact-checkers. Our approach employs controlled question generation to create precise queries that guide the verification process and retrieve relevant responses. This enhances the explanatory power of our model while ensuring data efficiency through clear, human-like inquiries. Empirical evaluations show that our framework significantly outperforms recent fact-checking baselines.

1 Introduction

The rise of online misinformation has become a pervasive and significant challenge, particularly in high-stakes contexts such as political elections and public health emergencies (Lewandowsky et al., 2020). The unrestricted spread of false narratives, misleading claims, and distorted statistics often devastate societal systems, causing political turmoil, economic instability, and decreased public trust in fundamental institutions. In response, various innovative fact-checking systems have emerged, offering scalable solutions to this growing misinformation epidemic (Saakyan et al., 2021; Sundriyal et al., 2022; Pan et al., 2023; Schlichtkrull et al., 2024). Despite significant advancements in fact-checking, one critical aspect remains underdeveloped: the verification of *quantitative claims* involving numbers, statistics, or temporal values.

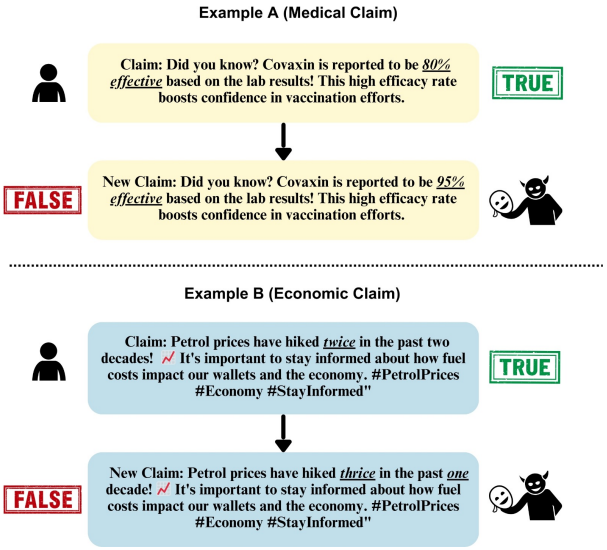

Figure 1: Representative examples of two quantitative claims and their subsequent modified *false* claims. The quantitative entities are underlined, and malicious users alter them to spread misinformation.

As highlighted by Sagara (2009), people are more likely to trust the information presented with numerical backing, even when it is false. This phenomenon, referred to as the *Illusion of Numeric Truth*, underscores how easily numeric data can influence perceptions. However, the very strength of numbers in shaping belief exposes a vulnerability to manipulation. A minor change in a numerical value can drastically alter a claim's meaning and public perception. For instance, a seemingly minor change in a percentage in Example A in Figure 1 can drastically alter the perceived effectiveness of a policy or product. In public health, a false claim stating that a vaccine is *'95% effective'* can lead to

a significantly different public perception than a claim stating it is *'80% effective.'* While both percentages indicate high efficacy, the increase from 80% to 95% may lead individuals to perceive the vaccine as significantly more reliable than it actually is, creating an unrealistic sense of security. Similarly, in economics, consider the claim *Petrol prices hiked twice in the past two decades.'* If it were stated as *thrice,'* difference in the numerical value could amplify concerns about rising costs. These discrepancies not only influence public opinion but can also impact policy decisions and consumer behaviour. The accuracy of such numerical claims is thus critical in shaping public understanding and reactions.

Despite the centrality of quantitative claims, verifying their accuracy remains a significant challenge for traditional fact-checking systems. While advancements in textual fact-checking are notable, the verification of numerical data is far more complex. Many existing fact-checking systems primarily rely on textual similarity or established fact databases (Thorne et al., 2018; Wang, 2017; Jiang et al., 2020), which fail to account for subtle changes in numerical information. This gap allows manipulated quantitative claims to bypass detection, leading to misinformed public perceptions. To address this challenge, we propose Q2FC, an innovative framework specifically designed to enhance the verification of quantitative claims, outperforming existing systems in effectiveness. Furthermore, current fact-checking datasets have not addressed the unique challenges posed by quantitative claims (Thorne et al., 2018; Wang, 2017; Jiang et al., 2020). The QuanTemp (Venktesh et al., 2024) dataset was recently introduced to handle numerical claims. While valuable, QuanTemp is limited in scope, covering only cardinal numeric data. To bridge this gap, we introduce QLAIM, a specialized fact-checking dataset that captures a broader range of quantitative claims. These include the date, time, percent, quantity, ordinal, and cardinal, enabling it to handle a broader spectrum of numerical data. For example, the claim *Alex placed first in the chess tournament.'* would be categorized as non-numerical by QuanTemp, as it lacks a cardinal Part of Speech tag. In contrast, QLAIM identifies *first* as an ordinal, recognizing it as a valid quantitative claim. QLAIM not only provides extensive coverage over numeric types but also provides nearly double the number of claims in QuanTemp.

Contributions. With this work, we offer the following contributions:[1]

- We highlight the unique challenges posed by the verification of quantitative claims.
- We create QLAIM, a comprehensive curated dataset focused on quantitative claims, with 33,422 fact-checked claims.
- We introduce Q2FC, a framework designed to tackle the verification of quantitative claims.

2 Related Work

Automated Fact-Checking. In recent years, automated fact-checking has seen significant progress, with models designed to detect misinformation across various domains. These systems typically involve claim detection (Gupta et al., 2021; Sundriyal et al., 2021), evidence retrieval (Aly et al., 2021), and veracity prediction (Pan et al., 2023; Schlichtkrull et al., 2024; Lee et al., 2020). The focus of most of these efforts has been on textual claims, verified against structured or unstructured data sources. Graph-based models (Zhou et al., 2019; Barnabò et al., 2023) have also been used to facilitate the reasoning over multiple pieces of evidence. Despite performance gains, these models struggle with explainability and require extensive training data. Recent studies indicate that LLMs can perform well and be dependable for verification tasks despite the possibility of hallucinations (Guan et al., 2024). Lee et al. (2020) demonstrated that the inherent knowledge of LLMs can be leveraged for fact verification. Previous research suggests that incorporating external information improves performance on reasoning-intensive tasks (Jiang et al., 2023; Yao et al., 2022). Recent advancements aim to simplify complex claims into manageable sub-questions to enhance evidence retrieval, showing promise in improving fact-checking accuracy, particularly for claims involving implicit reasoning or multiple verification steps.

[1]The source code and datasets are attached as appendices and will be available publicly upon acceptance of the paper.

Fact-Checking Datasets. Several fact-checking datasets exist, with the FEVER dataset being one of the most recognized, focusing on textual claims from Wikipedia. However, only about 10% of its claims involve numerical reasoning, a limitation shared by many datasets that often rely on synthetic or oversimplified claims. Datasets like TabFact and SciTab, which cover claims from Wikipedia tables and scientific contexts, still fall short in capturing the complexities of verifying numerical content in broader contexts. While real-world political datasets like ClaimDecomp, LIAR, and MultiFC include fact-checked claims, they don't specifically focus on numerical claims or the handling of statistical and temporal expressions. The QuanTemp dataset targets explicit numbers but lacks coverage of comparative, statistical, interval, and temporal aspects. In contrast, our dataset offers a more comprehensive approach, addressing a broader range of quantitative elements and providing a more nuanced understanding for developing advanced fact-checking models.

Question-Answering for Fact-Checking. Question-answering has emerged as a potential strategy for fact-checking. Yang et al. (2022) developed a model that doesn't require annotated question-answer datasets. Two recent datasets, Fan et al. (2020) and Chen et al. (2022), treat fact-checking as a question-answer task. However, these approaches face challenges: Fan et al. (2020) focuses on context rather than the full fact-checking process, while Ousidhoum et al. (2022) notes that many queries depend on external context, making them difficult to generate from the claim alone. Chen et al. (2022) attempted to ensure evidential sufficiency but faced challenges, including temporal leakage by relying on post-assertion publications. In contrast, Schlichtkrull et al. (2024) demonstrated that evidence reasoning can be efficiently modeled through question-answering, using human-generated questions and responses with supporting evidence. We propose automated quantitative entity-based question generation, which has advantages such as providing explanations beyond the facts, aligning with how humans analyze numerical data, and enabling the generation of controlled queries at scale.

3 Dataset

In this section, we outline the creation of the QLAIM dataset, designed to address the challenges of verifying quantitative claims through a multi-stage process ensuring domain diversity and suitability for automated fact-checking.

Data Collection. We initiate our data collection by sourcing claims from trusted fact-checking organizations through the ClaimReview Schema,[2] which is licensed under the Creative Commons Attribution-ShareAlike License (version 3.0). We adhere to the terms of this license. The initial collection encompasses a staggering 278,636 fact-checked claims spanning multiple languages and domains. However, to ensure the dataset's consistency and usability, we translate non-English claims into English using Google Translate and drop the languages not identified. One of the primary challenges in collecting these claims is the diversity of labelling conventions used by different fact-checking organizations. To address this, we standardize the labels for all claims to fall under three categories – True, False, or Not Enough Information. Claims with ambiguous labels or those lacking clear classifications were excluded. This standardization mirrors approaches in prior works, ensuring the dataset's compatibility with existing fact-checking pipelines. This refinement process yields a final set of 105,432 claims. After this filtering, we hone in on quantitative claims identification, which is detailed further in the following subsection.

Quantitative Entity Labelling. A key innovation in our dataset creation is the identification of quantitative segments within claims, termed *Quantitative Entity Labelling* (QEL). We test three tools for QEL: (a)*Regular Expression (RegEx)*, (b) *Named Entity Recognition (NER) tagging*, and (c) *Part of Speech (POS) tagging*. For the RegEx, we identify numerical values and choose a set of pre-defined terms often linked with quantitative statements, such as *increase, decrease, twice, double*, etc. For NER, we employ a pre-trained spaCy NER model to identify numeric types – Date, Time, Percent, Quantity, Ordinal, and Cardinal. Lastly, we use a BERT-based POS tagger (Hassan

[2] https://schema.org/ClaimReview

et al., 2022) to assess grammatical structure and identify Cardinal tags. Our human evaluation reveals that the RegEx and NER tagging approaches closely match human labels. The details of QEL human evaluation are given in Appendix A.2. Therefore, we incorporate both methods into our QEL process, merging their outputs to compile a comprehensive list of quantitative entities.

Data Statistics. Finally, our dataset contains 33,422 quantitative fact-checked claims. We partition it into an 80/10/10 split for training, development, and testing. Table 1 shows detailed statistics. Notably, most fact-checked claims are False, highlighting a larger tendency in the fact-checking arena, where fact-checkers frequently prioritize debunking misinformation over validating true claims. Detailed entity types and dataset samples are shown in Appendix A.1 and A.3.

Dataset	Train	Dev	Test
Number of claims	26737	3342	3343
Avg. claim length	128.49	129.34	130.33
Avg. questions per claim	1.55	1.54	1.53
Fact-check rating			
▷ False	21631	2704	2705
▷ True	4259	532	533
▷ NEI	847	106	105

Table 1: Data statistics of the QLAIM dataset. NEI denotes Not Englough Information.

4 Proposed Framework

Recently, Schlichtkrull et al. (2024) established that reasoning about evidence can be represented through questions and answers. Unlike them, who compose these questions manually, we use an automated approach to generate these human-like questions, focusing on the quantitative elements of the claim. We propose **Q2FC**, (Questioning Quantity for Fact-Checking), based on the perspective on assimilation of the correct questions and evidence. The overall framework is shown in Figure 2.

Q2FC's backbone comprises three sequential modules – controlled question generation, knowledge-grounded response generation, and veracity assessment. First, we denote the input claim as c, which is inherently quantitative. The process begins by generating a set of queries Q that specifically focus on the quantitative entities e present in c. This results in an ordered set of question-entity pairs $Q = \{(q_1, e_1), (q_2, e_2), \ldots, (q_m, e_m)\}$, where each query q_i corresponds to one quantitative entity e_i. Typical question-generation algorithms often struggle with creating queries for quantitative claims. Our methodology bridges this gap by ensuring that the generated questions are tailored to extract meaningful information about quantitative entities. Once the queries are defined, we use Large Language Models (LLMs) to retrieve accurate, contextually relevant responses. Finally, we compare the retrieved responses to the original claim c to determine its validity, indicating if the quantitative claim is supported. This systematic methodology enables a strong and efficient verification process for quantitative claims. The following subsections provide more information about each module.

Controlled Question Generation. Traditional methods of question generation often rely on supervised learning, which may lead to a lack of adaptability and insufficient contextual understanding. In contrast, we employ reward-based controlled question generation to enhance the generation of contextually relevant questions. We generate questions in a zero-shot manner. We then utilize a fine-tuned T5 model[3] to answer them by inputting both the original claim and the generated question. We then use Natural Language Inference (NLI) to determine whether the quantitative entity can be retrieved as the answer. The NLI scores are used as a reward. The Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) is employed to iteratively update the model based on these rewards derived from external evaluations. This dynamic learning approach allows the model to adapt and improve over time, effectively generating high-quality questions that are relevant and aligned with the provided context.

In each training iteration, the model generates a set of candidate questions for each claim based on prompts derived from the claim-quantitative entity pairs. The reward-based learning allows for continuous refinement of the model's parameters based on the rewards

[3] https://huggingface.co/MaRiOrOsSi/t5-base-finetuned-question-answering

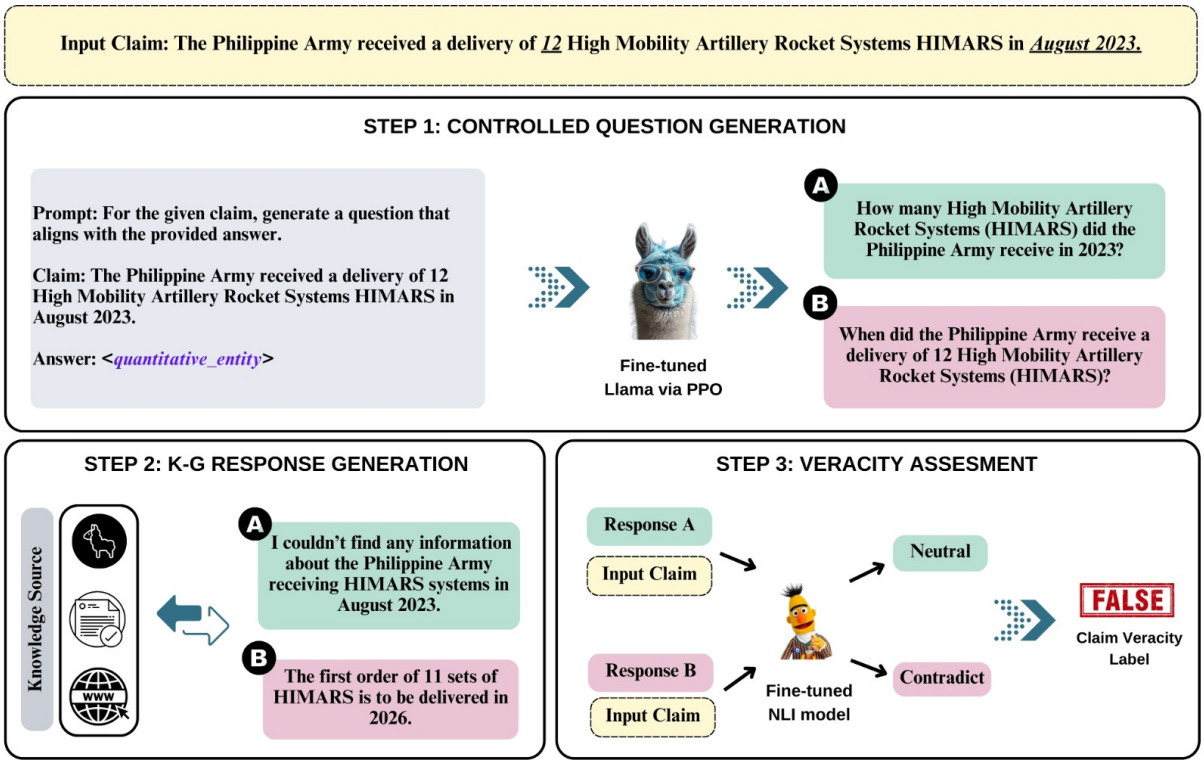

Figure 2: A schematic illustration of our approach, Q2FC, for a sample input claim, emphasizing quantitative components.(i) *Step 1*: Create questions for each quantitative entity. (ii) *Step 2*: Use a specific knowledge source to generate a response to the question. (iii) *Step 3*: Determine the final truthfulness label based on the alignment of the claims and the retrieved responses. 'K-G' stands for Knowledge-Grounded.

received, facilitating the generation of high-quality questions that are not only contextually aligned but also maximally informative.

Knowledge-Grounded Response Generation. After framing the questions, we extract responses from LLMs using three setups based on the available knowledge source (K).

Closed-Book Setup: In this setup, the model operates without external knowledge sources (K=ϕ), relying solely on its pre-trained internal knowledge to answer questions, highlighting its capabilities and limitations.

Limited-Book Setup: In this setup, K comprises a set of fact-checked articles that can explicitly support or refute the claim in question. Using the retrieval-augmented generation (RAG) (Lewis et al., 2020) framework, the model cross-references the claim with existing evidence to assess its veracity.

Open-Book Setup: In this setup, the model uses web-retrieved documents as its knowledge source. We create a corpus of the top five relevant documents for each claim using the Google Search API. The model indexes these documents via Facebook AI Similarity Search (FAISS) (Johnson et al., 2019) and employs the RAG to generate responses.

Veracity Assessment. After obtaining the responses, we use Natural Language Inference (NLI) to determine the claim's veracity. We determine whether the claim and generated responses logically support, contradict, or are neutral. We employ a Cross-Encoder model based on DistilRoBERTa, fine-tuned on the SNLI dataset (Bowman et al., 2015), to assess the logical relationships. If the responses support the claim, it is marked as *True*. If any response contradicts the claim, the claim is then marked *False*. Claims with unanswered questions are labelled as *Not Enough Information*.

5 Experiments and Results

Experimental Setup. For question generation, we experiment with five text-generation systems – BART (Lewis, 2019), T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2024), Gemma (Zoubarev et al., 2012), and Llama3

(Touvron et al., 2023). For the final veracity prediction, we benchmark our results against two of the latest fact-checking systems – AVeriTec (Schlichtkrull et al., 2024) and ProgramFC (Pan et al., 2023). All other implementation details are furnished in the Appendix A.6.

Performance Comparison. We present results for various experimental setups guided by the following research questions.

Which LLM is best for generating questions? Questions for a specific claim can vary in form due to different writing styles across news organizations. Capturing these variations is key to establishing a solid baseline for thorough analysis. While manual question generation is ideal, it is time-consuming. Thus, we explore how effectively existing LLMs can generate questions focused on quantitative entities using the following prompt:

```
For the answer <quantitative-
entity>, generate a question for
the given claim <claim>.
```

We generate multiple questions for each claim based on its quantitative entities using LLMs like BART, T5, FlanT5, Gemma, and Llama3. Some examples and the corresponding questions generated by the systems are shown in Appendix A.4. To evaluate the quality of these questions, 5 annotators[4] manually assess 75 randomly selected claim-question pairs across three dimensions: ▷ **Grammatical Correctness (GC):** This metric assesses the syntactic quality of the generated queries. ▷ **Factual Alignment (FA):** This assesses how well the questions match the factual content of the statements. ▷ **Relevance to the Quantitative Components (Rel):** This metric determines how closely the questions focus on the quantitative parts of the claim. As shown in Table 2, Llama3 outperforms all the other models across all dimensions and is used in subsequent experiments.

Which answering setup is most effective? To assess how diverse knowledge sources impact answering capabilities, we evaluate Q2FC across three settings, as outlined earlier. The results are shown in the bottom three rows of Table 3. Each setup offers different levels of access to external information, which is key for ensuring accurate quantitative answers. In the closed-book setup, the model relies on its pre-trained knowledge. While quick, it struggles with complex or time-sensitive queries, achieving an F1 score of 0.7043 and accuracy of 0.7586, the lowest of the three. The limited-book setup, with access to fact-checked articles, boosts accuracy and reliability, yielding the best performance: an F1 score of 0.7107 and accuracy of 0.7739. This is due to the use of verified sources. The open-book setup, leveraging online sources like Google searches, provides real-time information but varies in reliability. It achieves an F1 score of 0.7056 and accuracy of 0.7613, falling between the limited and closed-book setups.

How accurately do models predict veracity? Our Q2FC model consistently outperforms recent fact-checking systems, such as ProgramFC and AVeriTec, across all setups. ProgramFCachieves an F1 score of 0.7019 and accuracy of 0.7087, while AVeriTecscores 0.6398 in F1 and 0.6186 in accuracy. Even without external resources, the closed-book setup of Q2FC achieves an F1 score of 0.7043 and an accuracy of 0.7586, surpassing both systems. The limited-book setup, using fact-checked articles, delivers the best performance with an F1 score of 0.7107 and an accuracy of 0.7739, further outpacing both

Model	GC	FA	Rel
BART	0.4667	0.4445	0.5112
T5	4.8667	3.6445	3.6889
FlanT5	4.3442	3.1871	2.1542
Gemma	1.8889	1.1334	1.0000
Llama3	**5.0000**	**4.3556**	**4.2445**

Table 2: Average scores of manual evaluation for generated questions.

Model	F1	Acc
ProgramFC	0.7019	0.7087
AVeriTec	0.6398	0.6186
Q2FC *(closed-book)*	0.7043	0.7586
Q2FC *(limited-book)*	**0.7107**	**0.7739**
Q2FC *(open-book)*	0.7056	0.7613

Table 3: Experimental results for veracity labels. The last three rows show the results of our model with three distinct response generation setups.

[4]They were the advanced students (Masters and PhD) specialising in NLP, aged between 23 and 30.

systems. The open-book setup, using wider internet sources like Google searches, also surpasses ProgramFC, with an F1 score of 0.7056 and accuracy of 0.7613. Despite variability in source quality, it maintains better precision than earlier systems, demonstrating the advantage of diverse knowledge sources. Overall, Q2FC outperforms ProgramFC and AVeriTec in quantitative fact-checking, with the limited-book setup providing the highest precision and accuracy. This highlights the importance of reliable information sources in improving quantitative fact-checking systems.

Ablation Study. We present an ablation study to evaluate the impact of the Controlled Question Generation (CQG) module on the overall performance of our framework. The goal is to assess how the CQG module affects final veracity prediction, focusing on weighted F1 score and accuracy. In this study, we compare the performance of the Q2FC, with and without the CQG module. Specifically, we modify our retrieval process by using the original source claims to retrieve relevant evidence in Step 2, isolating the effect of the CQG module. In Step 3, we run the same NLI to ensure that any observed performance deviations are due to the CQG module. The results of the ablation study, presented in Table 4, show a clear improvement with the CQG module. Notably, including the CQG module improves both weighted F1 score and accuracy. The CQG module results in a 1.69% boost in weighted F1 score, while the accuracy improves by 1.52%. These findings demonstrate the CQG module's significant role in enhancing the framework's performance and its contribution to more accurate veracity prediction.

Configuration	F1	Acc
Without CQG	0.6989	0.7623
With CQG	0.7107	0.7739
Δ	1.69% ↑	1.52% ↑

Table 4: Impact of the Controlled Question Generation (CQG) module on Q2FC performance.

Qualitative Analysis. We investigate qualitatively how the quality and specificity of the generated answer are affected by varying accessibilities to external data. Through a comprehensive analysis, we identify key claim categories – real-time claims, historical claims, and predictive claims. We then examine how the three different setups (closed-book, limited-book, and open-book) respond to these claims, based on their respective knowledge access. A representative example from each claim category is provided in Table 5.

The closed-book system provides a generalised response for **real-time claims**, such as example 1 in Table 5, acknowledging its lack of real-time data and suggesting users consult other sources. The limited-book system improves by referencing historical data (71.66% of Florida in drought as of May 30), though it lacks real-time precision. The open-book system admits its inability to access real-time data but offers general drought patterns and directs users to reliable sources, providing more transparency but still lacking precision. For **historical claims**, like example 2, the closed-book system offers a generalized response, stating flu death figures are difficult to quantify due to seasonal variations but without concrete numbers. The limited-book system provides a more thorough response, citing reliable statistics from the WHO and CDC, enhancing factual accuracy. The open-book system also references general flu statistics but lacks the depth of the limited-book system, offering a more generalized response. For **predictive claims** such as example 3, the closed-book system provides cautious, generalized answers, highlighting its limitations in addressing future events. While acknowledging its inability to predict, the open-book system offers a general response referring to scientific consensus or existing regulations but lacks depth. The limited-book system offers corrective responses, such as discrediting the burqa ban claim with fact-checking sources and current regulations, providing a more accurate and thorough response.

Through this analysis, we observe that each setup handles different claims differently. The closed-book system is limited to generalized, cautious responses due to its lack of data access. The limited-book system offers more accurate, contextually rich responses, drawing on external knowledge, while the open-book system provides transparent answers with real-time data but still lacks access to specialized or future-specific information. The type of claim greatly influences each setup's effectiveness.

Example 1	**70 percent** of Florida is in drought today.
Question	What percentage of Florida is currently experiencing drought?
Closed-book	I'm not aware of any current severe weather conditions in Florida that is affecting 70% of the state. However, I can suggest some possible sources where you can find the most up-to-date information on current weather conditions in Florida.
Limited-book	According to the text, as of May 30, 71.66% of Florida was in a moderate to extreme drought. However, it's worth noting that this percentage has fluctuated throughout the years, and there have been periods of drought in Florida since 2000. Additionally, the text mentions that the return of the rainy season, which typically starts in June, should improve drought conditions.
Open-book	I don't have access to real-time information or specific data about current weather conditions in Florida.
Example 2	You lose **27000 people** to the common flu.
Question	How many people died from the common flu?
Closed-book	The exact number of deaths caused by the common flu (influenza) is difficult to determine, as it varies greatly from year to year and can depend on several factors such as the severity of the flu season, the effectiveness of the vaccine, and the population's age and health status. However, here are some general statistics...
Limited-book	Unfortunately, there isn't a definitive answer to this question, as the exact number of deaths from the common flu varies widely depending on several factors, including the specific strain of the flu, the population affected, and the availability of healthcare services. However, according to the World Health Organization (WHO), seasonal influenza (common flu) is responsible for approximately 3-5 million cases of severe illness and 290,000-645,000 hospitalizations annually worldwide.
Open-book	According to the World Health Organization (WHO), seasonal influenza epidemics result in approximately 3-5 million cases of severe illness and 290,000 to 646,000 hospitalizations worldwide each year. The exact number of deaths varies depending on several factors such as the severity of the flu season, vaccination rates, and healthcare infrastructure.
Example 3	From **May 20, 2025** headscarves and burqas will be banned in public swimming pools
Question	What is the date when headscarves and burqas will be banned in public swimming pools?
Closed-book	I'm not aware of any specific date for a global ban on headscarves and burqas in public swimming pools. However, some countries or local jurisdictions may have laws or regulations regarding swimwear or dress codes for public swimming pools.
Limited-book	The claim that 'From May 20, 2025 headscarves and burqas will be banned in public swimming pools' is incorrect according to the provided document from dpa fact checking. There is no evidence on the Internet that any swimming pool implemented a ban on headscarves or burkas on May 20, 2025. In Germany, there is no generally valid regulation that prohibits the wearing of headscarves or burqas in public spaces, including swimming pools. You only need to wear swimwear when swimming.
Open-book	I don't have access to specific information about future legislation or policy changes, including any potential bans on headscarves and burqas in public swimming pools.

Table 5: Qualitative analysis of the responses generated by our framework, Q2FC, using three different knowledge setups.

6 Conclusion

In an era driven by misinformation, the need for effective fact-checking approaches has become more pressing than ever. Misinformation, particularly in the form of misleading quantitative claims, poses significant challenges to public understanding and decision-making. Through this work, we addressed the critical gap in automated fact-checking systems, specifically focusing on the verification of quantitative claims, an area that is often overlooked in traditional methods. We developed a novel dataset, QLAIM, comprising more than 33k fact-checked quantitative claims. QLAIM encompasses a diverse range of quantitative claims including various numerical contexts – comparative, statistical, interval, temporal etc. We documented our data preparation process in detail, providing valuable insights for future research in this domain. We also proposed a framework, Q2FC, that introduces an question-answer based approach for fact-checking quantitative claims. We employed controlled question generation to create quantitative entity-based queries that drive the verification process. Empirical results showed that our technique outperforms existing baselines.

Limitations and Future Directions

While our work significantly advances quantitative fact-checking, we recognize and address probable limitations. First, during data gathering, we excluded claims containing images and videos. The decision was made based on the scope of our current investigation; nevertheless, we believe that including multimodal data could considerably improve the effectiveness of quantitative fact-checking. By incorporating images and videos, we may provide a more nuanced understanding of claims, as these kinds of media frequently provide context and connotations that textual claims alone may not. Furthermore, we understand the significance of expanding our activities beyond English. Considering misinformation knows no linguistic bounds, it is critical to create fact-checking systems that can work across multiple languages. By combining several languages, we can increase the global reliability and durability of quantitative fact-checking systems. This approach would allow for a more comprehensive understanding and study of assertions in various contexts, as well as successful engagement with varied populations. Language variety will assist us in addressing local misinformation issues and ensuring that our findings are relevant and applicable across diverse cultural contexts. Recognizing and resolving these limitations may allow us to improve the reliability and soundness of quantitative fact-checking systems in the future.

Ethics Statement and Social Impact

Data Bias. It is critical to consider the possibilities of biases in our dataset. Our data collection process includes acquiring fact-checked claims from a variety of fact-checking websites, each with its own set of editorial norms, procedures, and subjective interpretations. These aspects can introduce systemic biases that affect the overall evaluation of claims. However, we must acknowledge that we have no control over these biases.

Environmental Footprint. Large language models (LLMs) require a substantial amount of energy for training, which can contribute to global warming (Strubell et al., 2019). Our proposed approach for quantitative fact-checking leverages fine-tuning rather than training models from scratch, resulting in a significantly lower carbon footprint. Fine-tuning allows us to adapt pre-trained models to our specific needs with considerably less computational power and energy consumption. This not only minimizes our environmental impact but also enhances the efficiency of our fact-checking processes. It is important to note, however, that using LLMs for inference still consumes a considerable amount of energy. We seek to reduce this energy expenditure by exploring more energy-efficient techniques, such as pruning models, optimizing inference algorithms, and utilizing specialized hardware that minimizes power usage. By prioritizing sustainability in our approach, we aim to contribute positively to the ongoing conversation about the environmental implications of AI technologies.

Social Impact and Potential Use. Our model holds significant promise for the general public and can greatly benefit human fact-checkers by saving them time and resources. In an age where misinformation spreads rapidly across social media and other platforms, the need for reliable and efficient fact-checking has never been greater. By automating parts of the fact-checking process, our model can help identify and validate claims more swiftly, allowing fact-checkers to focus on more complex cases that require human judgment. This not only enhances the overall efficiency of fact-checking operations but also increases the accessibility of reliable information for the public.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13.

Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244.

Samuel Bowman, Gabor Angeli, Christopher Potts,

and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. 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 1090–1111.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188.

Sajjad Hassan, Durrani Nadir, Dalvi Fahim, Alam Firoj, Rafae Khan Abdul, and Xu Jia. 2022. Analyzing encoded concepts in transformer language models. In *North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL)*, NAACL '22, Seattle.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41.

Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracin, Panayiota Kendeou, Eryn J Newman, Gordon Pennycook, Ethan Porter, David G Rand, David N Rapp, et al. 2020. The debunking handbook 2020.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.

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. *Journal of machine learning research*, 21(140):1–67.

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 (Volume 1: Long Papers)*, pages 2116–2129.

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

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web.

Advances in Neural Information Processing Systems, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Megha Sundriyal, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, and Tanmoy Chakraborty. 2022. Document retrieval and claim verification to mitigate covid-19 misinformation. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 66–74.

Megha Sundriyal, Parantak Singh, Md Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Desyr: definition and syntactic representation based claim detection on the web. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1764–1773.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 650–660. Association for Computing Machinery (ACM).

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable fact-checking through question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956. IEEE.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.

Anton Zoubarev, Kelsey M Hamer, Kiran D Keshav, E Luke McCarthy, Joseph Roy C Santos, Thea Van Rossum, Cameron McDonald, Adam Hall, Xiang Wan, Raymond Lim, et al. 2012. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, 28(17):2272–2273.

A Appendix

A.1 Quantitative Entity Types

Table 8 gives an overview of different quantitative entity types in Named Entity Recognition (NER) that we used for our dataset. These entity types cover varied categories, such as DATE, which records absolute or relative references to dates and periods, and TIME, which signifies time spans smaller than one day, like particular hours or minutes. The PERCENT type recognizes percentages, while MONEY notes down monetary values with their related units. QUANTITY type is about measurable amounts, like weight or distance. Then, ORDINAL means numbers that show a certain rank or order, such as 'first' or 'second.' Lastly, CARDINAL covers numerical values not included in the prior categories and acts as broad numeric identifiers. Using the spaCy library for NER tagging to extract these categories allows the model to properly differentiate and sort out numeral entities present in the text, bettering its capacity to process and comprehend quantity-based data across different situations.

We present a detailed breakdown of the statistics of these quantitative entities within QLAIM, in Table 6. It is important to note that a claim may contain multiple types of quantitative entities.

Entity Type	Train	Dev	Test
DATE	10794	1407	1323
TIME	921	126	124
PERCENT	494	56	69
MONEY	705	90	71
QUANTITY	468	63	60
ORDINAL	1614	196	202
CARDINAL	10828	1336	1322

Table 6: Statistics of the quantitative entity types within QLAIM.

A.2 Human Evaluation of QEL

To evaluate the QEL approaches, we conducted a human assessment with three annotators (author and two undergraduate students working in NLP, aged 22-27 years) manually identifying quantitive entities in 100 random samples from our dataset. We compare the human-labelled entities with those generated by the tools using the Fuzzy and Jaccard scores. The Jaccard and Fuzzy scores are commonly used similarity metrics. Jaccard Similarity[5] is a set-based metric that calculates the ratio of the intersection of two sets to their union, which measures the overlap between the sets. We calculated the Jaccard score by dividing the intersection of the human-labelled entities and those produced by the tools by the union of these two sets. In contrast, the Fuzzy Score[6] measures the similarity between two strings, accounting for approximate matches such as minor typographical errors or variations in word order. The Fuzzy score uses Levenshtein Distance, which computes the number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. For each entity in the tool-generated set, we compare it to every entity in the human-labeled set (ground truth) and calculate the Fuzzy Score.

Metric	RegEx	NER	POS	RegEx ∪ NER
Fuzzy	0.7723	0.8460	0.6466	**0.9168**
Jaccard	0.4272	0.6843	0.2140	**0.7929**

Table 7: Human evaluation results of quantitative entity labelling approaches.

The results shown in Table 7 reveal that the RegEx and NER tagging approaches closely match human labels. Therefore, we incorporate both methods into our QEL process, merging their outputs to compile a comprehensive list of quantitative entities. Claims without these entities are excluded from the final dataset.

A.3 Data Analysis

Table 9 shows examples from our QLAIM dataset. It shows how we pull out numbers and quantities from different claims. The Claim column, in Table 9, contains statements involving numerical or temporal details, while the Quantitative Entity column highlights the extracted numerical entities associated with each claim. Note that there can be multiple entities in one claim. As an example, in the first example, the quantity *[50x, September]* represents how much change happened and when it happened. These are vital to understand the factual basis of the claims.

[5] https://en.wikipedia.org/wiki/Jaccard_index

[6] https://github.com/seatgeek/thefuzz

Entity Type	Description	Example
DATE	Absolute or relative dates or periods	Covid-19 was announced as 'Pandemic' on 29 February 2020.
TIME	Times smaller than a day	The meeting is scheduled for 3 pm.
PERCENT	Percentage (including "%")	About 50% of the population voted in the last assembly elections.
MONEY	Monetary values, including unit	The book's price, is $19.99.
QUANTITY	Measurements such as weight or distance	The package weighs 10 kg.
ORDINAL	"first," "second," etc.	She finished in first in the race.
CARDINAL	Numerals that do not fall under another type	There are 25 students in the class.

Table 8: Broad overview of numeric entity types, along with their descriptions and examples. We use spaCy for the NER tagging.

Claim	Quantitative Entity
Virus levels are now 50x higher among secondary school pupils than they were in September	*[50x, September]*
A ghost bus filled with FBI informants dressed as Trump supporters deployed onto our Capitol on January 6th	*[January 6th]*
Over 800 pounds of fentanyl were seized at our Southern Border in October 2023. This is Biden 2019s Border Crisis	*[800 pounds, October 2023]*

Table 9: Examples from our dataset, QLAIM, along with their quantitative entities.

A.4 Question Generation

We assess the ability of various language models –BART, T5, FlanT5, Gemma, and Llama3 – to generate questions about numerical values within provided claims. Some samples are presented in Table 11. Overall, T5 and Llama3 excel in inquiring about numerical components such as dates and numbers. Nevertheless, disparities occur among models. While FlanT5 occasionally misinterprets claims, resulting in questions that are out of place, BART has problems with coherence and occasionally produces nonsensical content. Despite being structured, Gemma frequently uses excessive amounts of words and provides explanations instead of asking a direct question. T5 and Llama3 are the most promising, but further fine-tuning is necessary to prevent irrelevant or off-topic question formation.

A.5 Model Performance Across Entity Types

The Q2FC demonstrates strong performance for DATE and TIME entities, with the highest F1 and accuracy scores, as shown in Table 10. However, because of their contextual unpredictability, MONEY and PERCENT entities perform worse. The modest performance of the ORDINAL and CARDINAL entities indicates that claims based on counting and numerical order are handled rather well. Additionally, QUANTITY entities attain a respectable level of precision, suggesting that the model can accurately process numerical measurements.

Type	F1	Accuracy
DATE	0.7021	0.7725
TIME	0.7609	0.8065
PERCENT	0.4970	0.5942
MONEY	0.4294	0.5493
QUANTITY	0.6205	0.7333
ORDINAL	0.6859	0.7376
CARDINAL	0.6605	0.7428

Table 10: Performance of our model across different quantitative entity types.

A.6 Implementation Details

Our experiments utilize the Llama3.18B variant throughout, with a maximum input length set to 128 tokens. The Llama model is configured with a temperature of 0.2 and a repetition penalty of 1.1. For BART, T5, and FlanT5, we use the large variants. We load these models via the *Hugging Face library*. For Proximal Policy Optimization (PPO) training, we leverage Python's *trl* library.[7] We use a learning rate of 1.4e5 and a batch size of 8. Our PPO model is

[7] https://huggingface.co/docs/trl/en/index

trained for 10 epochs. We utilize weighted F1 and accuracy from the *sklearn library* to compare our proposed framework's performance with these systems. All experiments are carried out on a system running NVIDIA Driver version 525.147.05 and CUDA version 12.0. The GPU used is an NVIDIA A100 with 80GB of RAM.

Claim	Over **30000** scientists worldwide declare climate change is a hoax.
BART	Is climate change a real hoax according to scientists?
T5	How many scientists believe climate change is a hoax?
FlanT5	What is the name of the group that declared climate change a hoax?
Gemma	Can you name the scientists who are against the climate change?
Llama3	What is the approximate number of scientists who have allegedly declared climate change a hoax?
Claim	Smoking in cars with children is illegal starting **January 30, 2018**
BART	Smoking in cars with children is illegal starting January 30 2018OSE tv tvOSE subversive subversive subversiveptions...
T5	When will it become illegal to smoke in cars with children?
FlanT5	What is the name of the group that declared climate change a hoax?
Gemma	Context: Smoking in cars with children is illegal starting January 30 2018 Answer: January 30 2018 Explanation: The question is generated by using the context and the answer.
Llama3	What is the date when smoking in cars with children became illegal?
Claim	A **1922** newspaper article reported that radical change in climatic conditions was melting Arctic ice and disrupting wildlife.
BART	A 1922 newspaper article reported that radical change in climatic conditions was melting Arctic ice and disrupting wildlife...
T5	What year was the article published?
FlanT5	What year was the article published about climatic conditions and melting Arctic ice?
Gemma	A 1922 newspaper article reported that radical change in climatic conditions was melting Arctic ice and disrupting wildlife Answer: 1922 newspaper article reported that radical change in climatic conditions was melting Arctic ice and disrupting wildlife...
Llama3	What is the date when smoking in cars with children became illegal?

Table 11: Examples of quantitative claims and corresponding generated questions from various models. The quantitative entities are highlighted in **bold**.