Towards Improving Retrieval for the Verification of Natural Numerical Claims

Deepali Prabhu

Towards Improving Retrieval for the Verification of Natural Numerical Claims

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Deepali Prabhu
born in Bengaluru, India

Software Engineering Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Towards Improving Retrieval for the Verification of Natural Numerical Claims

Author: Deepali Prabhu
Student id: 5732166
Email: D.Prabhu@student.tudelft.nl

Abstract

Verification of numerical claims is critical as they tend to be more believable despite being fake and have previously demonstrated the potential to cause catastrophic impacts on society. While there currently exist several automatic fact verification pipelines, only a handful focus on natural numerical claims. A typical human fact-checker first retrieves relevant evidence addressing the different numerical aspects of the claim and then reasons about them to predict the veracity of the claim. Hence, the retrieval thought process of a human fact-checker is a crucial skill that forms the foundation of the verification process. Emulating a real-world setting is essential to aid in the development of automated methods that encompass such skills. Hence, we introduce QUANTEMP++: a dataset consisting of natural numerical claims, an open domain corpus, and the corresponding evidence relevance and veracity labels. Given this dataset, we also aim to characterize the retrieval performance of key query planning paradigms, especially those of decomposition as they have shown promising results in other tasks. Finally, we observe their effect on the outcome of the verification pipeline and draw insights.

Thesis Committee:

Chair:	Prof. Dr. Avishek Anand, Web information Systems, TU Delft
University supervisor:	Dr. Venktesh Viswanathan, Web information Systems, TU Delft
Committee Member:	Prof. Dr. Pradeep Murukannaiah, Interactive Intelligence, TU Delft

Preface

The past two years of my life have been a roller coaster. As I reflect on this period, I am deeply indebted to a few individuals who cheered me on during the highs and motivated me through the lows.

Firstly, I would like to thank my thesis advisor, professor Avishek Anand, for providing me this incredible opportunity to work on a subject that aligns so closely with my passions and for always dedicating the time needed to point me in the right direction. I am also indebted to my daily supervisor Venktesh Viswanathan for supporting my ideas and having patience through my ocean of questions and debugging sessions. Additionally, I would like to thank everyone in the Web Information Research Group for their valuable feedback and for making me feel included in their research community.

I would also like to express my deepest gratitude to Varun and Rahul for their steadfast support, their attentive listening to my concerns, and their invaluable advice. I would like to thank my seniors, Sayak and Sreeparna, for sharing their experiences and helping me understand and prepare for the realities of completing a thesis. Finally, I would like to extend my heartfelt thanks to the constants in my life over the past two years: Jai, Vishakha, and Kalaivanan. From the countless heartwarming meals after setbacks to the calming and relaxing walk-and-talks that helped me unwind, I could not have come this far without their unwavering support. I am deeply grateful to Ahandeep for standing by my side over these past two years, showing incredible patience and understanding. I couldn't have achieved this without his unconditional support.

My vote of thanks would be incomplete without expressing my gratitude to the most important people in my life, my Amma and Papa. You have always supported me, sacrificing your own comfort to see me succeed. Your dedication and selflessness have been both an inspiration and a driving force throughout this journey.

Over the course of two years, I believe I have grown immensely, through late nights of problem-solving to collaborative efforts. I am grateful to everyone who has played a part in shaping this work.

Deepali Prabhu
Delft, the Netherlands

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Terminology

A *natural claim* is a claim made in everyday natural language usage as opposed to those generated synthetically by a computer. We define a *natural numeric claim* as any *natural claim* that contains numerical information in the form of quantities, dates, statistics, or comparisons. This numerical information may be explicit as in the claim "There were 20 children in the playground" or implicit as in the claim "The crime rate has reduced over the past years".

1.2 Motivation

With the growth of digital tools for disseminating information, misinformation and disinformation have also increased tremendously in recent years. Although manual verification is performed by journalists in news organizations and on dedicated websites, the sheer volume of claims generated calls for automated approaches, representing a key objective of computational journalism [12] .

The case of real-world numerical claims. Numerical claims are more important to verify as by the Illusion-of-Numeric-Truth effect [54], people tend to believe claims grounded in numbers more even if they are false. Such claims can have various negative impacts on society. An example of this is the case of Purdue Pharma, where unsubstantiated, incomplete claims were made to market their drug which is believed to be the kickstarter of the Opioid pandemic in America, resulting in the loss of over 500,000 lives to date[66]. Verification of numerical claims requires numerical reasoning ability which involves understanding numeric patterns and mathematical concepts like arithmetic, numerical estimation, and data interpretation. While several works focus on automated verification of natural claims, only a handful[43, 59, 68, 63, 58, 26] focus on claims that require numerical reasoning. However, these techniques only focus on the detection of numerical claims[58, 59] or are restricted to specific statistical properties[63, 68].

The need for a realistic fact-checking dataset. A typical human fact-checker utilizes various skills to verify a natural numerical claim. A given numerical claim can have different explicit and implicit information needs that first needs to be addressed. Therefore, relevant evidence is gathered from various sources, including the open web. This evidence is then used to reason

Example: Claim from QUANTEMP++

Claim: `Prime Minister Narendra Modi breached the election protocol by addressing a rally in Howrah on April 6.`

***Ideal Retrieval*:**

- ...From cooch behar ...Howrah ...election rally on April 6, pm Modi will address...
- ...Election Silence Period...India (48 to 24 hours in advance of polling day and on polling day)...
- ...General elections were held in India in seven phases from 11 April to 19 May 2019...

***Retrieval Thought Process*:** The given numerical claim requires multiple aspects to be addressed. Firstly, evidence proving that PM Modi did indeed rally on April 6 needs to be fetched. Then, the election protocol breached i.e. information on the election silence period and the schedule of elections (polling) are required to be fetched. Given this information, the downstream verifier can then reason about the numerical aspects to form the verdict.

Figure 1.1: Example numerical claim from QUANTEMP++, requiring to recognize and fetch explicit and implicit evidence.

about and determine the veracity of a natural numerical claim. The retrieval thought process of a human fact-checker is a crucial skill that forms the foundation of the verification process (See Figure.1.1). Emulating a real-world setting is essential to aid in the development of automated methods that encompass such skills.

Current datasets mainly feature synthetic claims with lexical biases [44, 10], while those with natural claims often suffer from gold or temporal leakage[56] or lack significant numerical claims (see Table 1.1). The presence of leaked fact verification articles in the corpus trivializes retrieval, while including evidence published after the claim makes the corpus unrealistic. Consequently, any automatic method designed under these conditions is likely to fail during real-time deployment. Hence, our goal is to provide a realistic dataset that addresses all these gaps, aiding the right development of automated methods for verifying natural numerical claims.

Forming a large-scale dataset to address the above gaps through crowdsourcing can be challenging as getting manual annotations from expert fact-checkers is expensive[15] and can often contain biases[44]. Therefore, we employ weak supervision to extend QuanTemp[65] and create **QUANTEMP++**: a dataset consisting of about 15k natural numerical claims, an open domain corpus consisting of 165.7k records, and the corresponding evidence relevance and veracity labels of the claims. The data creation pipeline incorporates claim decomposition to emulate the thought process of human fact-checkers in retrieving evidence, which helps in realistically evaluating and addressing retrieval bottlenecks in automated fact-checking methods.

Paper	#Claims	Claim Src	Natural?	Evidence Src	No Leakage		Numerical*
					Gold	Temporal	
Thorne and Vlachos[63]	7k	KB	✗	KB	NA	NA	✗
Vlachos and Ridel[68]	7k	KB	✗	KB	NA	NA	✓
FavIQ[44]	188k	QA	✗	Wikipedia	NA	NA	✗
VitaminC[57]	326k	Wikipedia	✗	Wikipedia	NA	NA	✗
HOVER[27]	26k	QA	✗	Wikipedia	NA	NA	✗
SciFact[71]	1.4k	Science Articles	✗	Science Articles	NA	NA	✗
SciFactOpen[72]	279	Science Articles	✗	Science Articles	NA	NA	✗
WICE[30]	2k	Wikipedia	✗	Wikipedia	NA	NA	✗
FEVER[64]	185k	Wikipedia	✗	Wikipedia	NA	NA	✗
FEVEROUS[4]	87k	Wikipedia	✗	Wikipedia	NA	NA	✗
MultiFC[5]	36k	FCW	✓	Open Web	✗	✗	✗
ClaimDecomp[9]	1.2k	FCW	✓	Fact Check Articles	NA	NA	✗
FinFact[47]	3.4k	FCW	✓	Open Web	✗	✗	✗
WatClaimCheck[33]	34k	FCW	✓	Open Web	✓	✓	✗
LIAR[75]	13k	FCW	✓	NA	NA	NA	✗
QABriefs[16]	8.8k	FCW	✓	Open web	✓	✗	✗
AviriTec[56]	4.5k	FCW	✓	Open Web	✓	✓	✗
Jandhagi and Pujara[26]	500	News Articles	✓	Select Datasets	✓	✓	✓
COVIDFACT[53]	4.1k	Reddit	✓	Open Web	✓	✓	✗
QuanTemp[65]	15k	FCW	✓	Open Web	✓	✗	✓
QUANTEMP++(OURS)	15k	FCW	✓	Open Web	✓	✓	✓

Table 1.1: Comparison of QUANTEMP++ to other related fact checking datasets. Leakage refers to either the presence of fact-checking articles (Gold) in the evidence corpus or the inclusion of evidence items published after the corresponding claim (Temporal). We also judge if a significant number of Numerical* claims are present in the dataset. In the table, FCW refers to Fact-Checking Websites. The number of claims listed in the table are approximate figures to indicate the scale.

The role of query planning. A typical automatic fact verification pipeline consists of three stages: claim detection, evidence retrieval, and veracity prediction. Although many effective automatic fact-verification pipelines have been proposed in previous research, many of these approaches highlight bottlenecks encountered during the retrieval stage.[33, 44, 71]. Within retrieval, query planning plays an important role as it defines the information needs to be met in order to verify a claim, forming the entry point of retrieval. Since numerical claims encompass various implicit and explicit numerical aspects (See Figure.1.1), the query planning method utilized can have a significant impact on the retrieval of evidence and hence on the downstream veracity classification performance.

Popular fact verification pipelines like MultiFC[5], FEVER[64], FEVEROUS[4], SciFact[71], and WatClaimCheck[33] utilize only the claim as the query to fetch the required evidence. However, following this strategy has previously been shown to reduce recall[27, 44]. Several techniques have been used previously to improve this aspect, one of the main ones being decomposition. Works like ClaimDecomp[9], WICE[30], QuanTemp[65], and ProgramFC[43] show the benefits of decomposition on the downstream performance of the verification of claims. However, they do not characterize the retrieval performance in a realistic open-domain setting for natural numerical claims. Additionally, in previous works, we see there is a lack of a principled

aggregation method to combine the retrieval results from the decomposed queries of a claim. In our work, we evaluate the retrieval performance of key query planning methods with a principled aggregation method to combine results across sub-queries. Finally, we analyze their downstream impact and draw insights.

In summary, we set out to address the following research questions:

1. Does query decomposition help retrieve quality evidence from the web for the verification of natural numerical claims?

2. How do existing query planning methods perform in terms of retrieval of relevant evidence snippets to verify numerical claims?

3. What is the downstream impact of these query planning methods on the task of verification of numerical claims?

This translates to the following contributions:

1. A new dataset, QUANTEMP++, to evaluate evidence retrieval for fact-checking natural numerical claims in an open-domain realistic setting.

2. A comprehensive evaluation of key query planning methods for evidence retrieval to support fact-checking numerical claims.

3. A comprehensive evaluation of the downstream impact of these key query planning methods at the task of fact-checking numerical claims.

1.2.1 Thesis Outline

The remainder of this thesis is structured as follows: Chapter.2 identifies the research gaps in related work that this study aims to address. Chapter.3 describes the data creation pipeline and the methodology used to evaluate key query planning methods. Chapter.4 details the experiments conducted and their implementation specifics. Chapter.5 discusses the results of these experiments and provides insights. Finally, Chapter.6 concludes the thesis, highlighting limitations and suggesting directions for future research.

Chapter 2

Related Work

Automatic Fact-checking is a well-explored task, first introduced formally by Vlachos and Riedel [67] in 2014. A fact verification pipeline mainly consists of three stages- claim detection, evidence retrieval, and claim verification. A few recent works also include an additional stage for justification production [52]. This paper exclusively concentrates on the evidence retrieval stage, crafting a natural numerical dataset to capture its challenges and, in turn, bolstering the numerical reasoning capabilities essential for successful fact verification. In this section, we delve into current datasets concerning natural claims, natural numerical claims, the challenges they pose, and the methods devised to improve retrieval for numerical reasoning.

2.1 Datasets

Since the formulation of the task, various datasets consisting of natural claims have been released in several domains such as politics[9, 75, 3, 42], finance[59], climate[14], health[37], scientific research[71, 69, 77], and social media[41, 13]. These datasets vary by the sources from which these claims are gathered and how they establish evidence labels, if indeed they do. A significant portion of these datasets collect claims sourced from Wikipedia[64, 4, 57, 30]. FEVER[64], a popular dataset in the domain of fact verification, constructs a dataset of claims by employing crowd workers to rephrase phrases from Wikipedia articles. FEVEROUS[4] takes it one step further by forming longer and more complex claims using the FEVER[64] dataset and its associated tables from Wikipedia. While these techniques are effective in generating natural datasets of claims that require verification, these datasets are prone to be infected by lexical bias. For example, studies by Park et al. [44] found that the top bi-grams in refute claims of the FEVER dataset [64] contain negative expressions, e.g., “is only”, “incapable of”, “did not” and hence are significantly lexically biased.

To alleviate this, VitaminC[57] leverages Wikipedia revisions to generate challenging examples in which a claim is paired with contexts that are lexically similar, yet factually opposing. WICE[30] aims to accomplish the aforementioned goal through the automation of claim generation by using Wikipedia citations. Scifact[71] employs a similar strategy but with scientific research articles as their source. Although this method is efficient in generating extensive datasets

for claims, the synthetic nature of the data makes them unrepresentative of naturally occurring claims.

2.1.1 Natural Claims

To form a dataset of natural claims, Hover[27] and FaVIQ[44] attempt to generate claims using question-answering datasets. HoVer[27] employs crowd workers to form long-range natural claims using question-answer pairs from HotpotQA [81] that demand multi-hop reasoning for verification. Since employing expert crowd workers to generate claims is expensive, FaVIQ [44] automates this process by training a T5 model [49] to convert question-answer pairs from AmbigQA [39] to natural claims. A more straightforward approach is taken by various methods, where fact-checking websites are crawled for natural claims [22]. While datasets such as LIAR[75], Covid_Fact [53], PUBHEALTH [36] are domain-specific, MultiFC[5] and WatClaimCheck[33] present datasets that are open domain.

The aforementioned studies employ various methods for collecting gold evidence, including Search APIs, weak supervision, citations, and the utilization of review articles. WatClaimCheck[33] utilizes review articles accompanying claims in fact-checking websites and the links in them to form the evidence for claims. However such datasets created by professional fact-checkers are expensive and small-scale. Scifact[71] and FEVEROUS[4] use citations to form the evidence set, however, these approaches do not operate within open-domain retrieval settings. To alleviate the above disadvantages, MultiFC[5] uses the Google search API to collect evidence articles for each claim. Such a dynamic search retrieves evidence on the fly from the open web, making reproducibility impossible. Additionally, MultiFC[5] does not address the prevention of gold evidence leakage and temporal leakage[56] during its search. Scifact-Open[72], an extension of Scifact[71] designed for open-domain settings, and FaVIQ[44] both employ weak supervision techniques to fetch evidence labels. Scifact-Open[72] utilizes simple BM25 retrievers with a pooling strategy, while FaVIQ[44] leverages Dense Passage Retrieval[32]. It is important to note that these evidence labels can be noisy as underlying systems are not perfect.

2.1.2 Natural Numerical Claims

These aforementioned datasets demand several abilities such as including multi-hop reasoning, retrieval from heterogeneous data, and processing long claims. However, very few have significant samples that require numerical reasoning. For example, the authors of FEVEROUS [4] observe that although merely around 10% of the claims require numerical reasoning, the annotators considered these particular claims to be especially challenging. Addressing numerical claims can be convoluted, necessitating a range of skills, including the extraction of numerical keywords and symbols, comparison, recognition of recurrent patterns, trend detection, utilization of higher-order functions, and interpretation of intervals over time, among others. Very Few works focus on generating and dealing with numerical natural claims.

The method proposed by Shah et al. [59] and CONCORD [58] detect natural numerical claims from financial reports and academic papers respectively. While they use weak supervision to automate claim detection, they do not propose methods to form labels for retrieval and verification. Similarly, FinFact[47] presents a benchmark dataset for multimodal fact-checking

within the financial domain, including professional fact-checker annotations and justifications. The methods proposed by Thorne and Vlachos [63] and Vlachos and Riedel [68] present datasets for natural numerical claims along with evidence and verification labels but only focus on simple claims or work with limited statistical properties. Additionally, both are limited to retrieving evidence from knowledge graphs. Jandaghi and Pujara [26] identify numerical claims from news articles by using structure-based detection. Additionally, these claims are aligned with evidence by extracting indicator and trend entities from claims and matching them with those of the evidence. Lastly, QuanTemp[65] introduces a comprehensive dataset containing purely natural quantitative and temporal claims. However, while a corpus of Google Search snippets is provided, the dataset lacks relevance labels and contains temporal leakage.

2.2 Evidence Retrieval

Retrieval of evidence forms a critical component in the fact verification pipeline. Several works like WatClaimCheck[33] show that retrieval serves as a performance bottleneck in the pipeline. Studies conducted by Park et al. [44] demonstrate that the most common error occurring in the fact verification pipeline is during retrieval. Experiments conducted by Scifact[71] demonstrate that given oracle retrieval results, the fact verification pipeline demonstrates a huge gain in performance. Additionally, a strong retrieval model can facilitate human-in-the-loop verification systems. Fan et al. [16] show that a strong evidence retriever increases the accuracy of crowd workers by 10% while slightly decreasing the time taken for the task of fact verification.

2.2.1 Query Planning

Query planning plays an important role in retrieval as it explicitly conveys the information needs of the claim. Most popular methods like MultiFC[5], FEVER[64], FEVEROUS[4], SciFact[71], and WatClaimCheck[33] utilize only the claim as the query to fetch required evidence. However, following this strategy has previously been shown to reduce recall, especially when there are multiple aspects contained in the claim[27, 44]. Numerical claims encompass various aspects that must be addressed, as the downstream veracity classifier processes these multiple pieces of evidence to reason numerically or temporally, perform calculations, or make comparisons. Several techniques have been used previously to improve retrieval, one of the main ones being decomposition.

Decomposition

Decomposition works to address different aspects of a claim explicitly in the retrieval pipeline. Previous works use two main paradigms to decompose claims into sub-queries.

Fine-Tuning: The first paradigm involves crowdsourcing and utilizing human annotations to fine-tune smaller language models, as seen in ClaimDecomp[9] and QABriefs[16]. While both ClaimDecomp and QABriefs[16] show that decomposition provides superior downstream performance, their retrieval setting is not realistic. ClaimDecomp [9] operates in a proof-of-concept setting, whereas QABriefs [16] dynamically queries the web to retrieve top documents for each claim, without attempting to prevent temporal or gold leakage.

Prompting: A major disadvantage of the above paradigm is the reliance on expert annotations to decompose claims. This approach requires a substantial amount of data for effective performance, which is not always readily accessible. Hence, more recent methods utilize prompting strategies to overcome this challenge. Works like ProgramFC[43], QuanTemp[65], and WICE[30] use few shot learning to generate the decomposed questions for a claim. However, ProgramFC[43] and WICE[30] operate on synthetic claims generated from Wikipedia, and QuanTemp[65] dynamically queries the web without attempting to prevent temporal leakage. Additionally, generating decompositions through prompting can suffer from hallucinations when there is a lack of background information, leading to errors downstream[85]. To overcome this, Zhang and Gao [85] and Wang and Shu [74] utilize a multi-step prompting strategy augmented with a dynamic web search to provide context to facilitate query decomposition. While this reduces hallucination and improves overall decomposition quality, its multi-prompt nature makes it extremely expensive.

Knowledge Distillation: Recent works have shown that knowledge distillation from large models to smaller models is an effective strategy to reduce inference costs while maintaining performance [62, 11, 45, 21]. Specifically, FlanT5 [38] has shown to outperform zero shot LLMs like GPT-3.5[7] at specific in-domain tasks such as summarization[18]. Additionally, Wu et al. [79] has demonstrated that problem decomposition tasks can be more easily distilled into smaller models such as Vicuna from larger LLMs like GPT-3.5 for mathematical reasoning and QA tasks, thereby reducing inference costs while maintaining performance. Hence, in our work, we try to adapt this strategy to our task and evaluate this mode of query planning.

While the above methods show decomposition can be helpful, we observe that the above methods either don't evaluate retrieval and assume access to oracle evidence[9, 30] or they don't have a principled method to combine the retrieved results from each sub-query[65, 43, 16]. V et al. [65] and Fan et al. [16] use the top result per query to form the final evidence set. However, using these methods can miss out on retrieving important aspects as each sub-query itself can be ambiguous requiring multiple documents to cover each perspective. Additionally, the size of the resulting evidence set varies by the number of sub-queries generated, making comparison across methods difficult. Previous work in information fusion[70, 86, 84] has shown us effective techniques to combine results from different queries such as CombMAX and CombSUM. These methods take the relevance scores assigned to all the documents by the retriever for each query and assign the final score to a document using a maximum or sum function over them to form the final score for each document. In our work, we evaluate the utility of these fusion techniques in the domain of query decomposition for the verification of natural numerical claims.

2.3 Related Work Relevancy

In recent works, we observe that while there are several datasets for the verification of claims, only a handful of them focus on natural numerical claims(See Figure.1.1). Additionally, the existing datasets with natural claims often do not provide relevance labels for an open-domain retrieval setting or their retrieval stage corpus has gold evidence or temporal leakage. The few natural datasets that evaluate retrieval either use only the claim to retrieve evidence, which is said to reduce recall, or use decomposition in a closed domain setting with no principled way

of aggregating retrieval results across sub-queries. Since numerical claims require an effective query-planning method to retrieve evidence that addresses multiple possible explicit and implicit aspects, a thorough and principled evaluation of query-planning methods is necessary to understand their pros and cons. In our work, we aim to address these research gaps by providing a dataset for natural numerical claims with evidence-relevance labels in an open-domain setting without gold evidence or temporal leakage. Furthermore, using this dataset, we evaluate the retrieval performance of key query planning methods with a principled aggregation method to combine results across sub-queries. Finally, we analyze their downstream impact and draw insights.

Chapter 3

Methodology

3.1 Dataset Creation

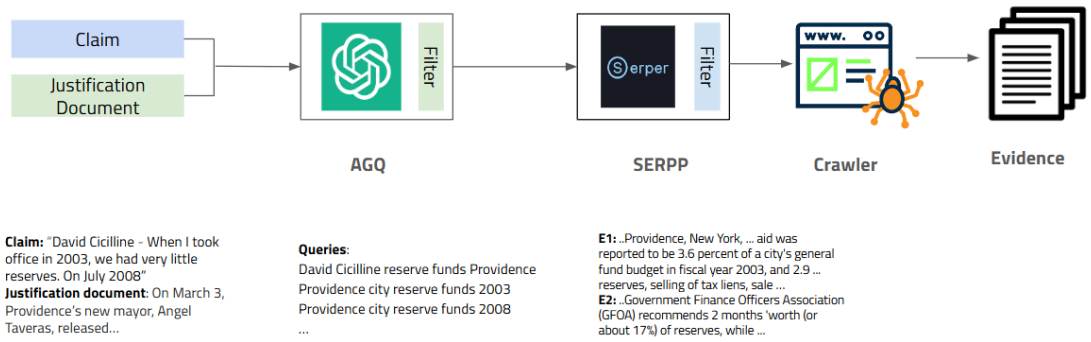


Figure 3.1: Data creation pipeline of QUANTEMP++

The first contribution of this work is to present a new dataset to evaluate evidence retrieval for fact-checking claims with quantitative and temporal expressions. This high-quality dataset can then be employed to evaluate existing retrieval methods, specifically query planning. Hence, the main goal of this dataset is to capture the reasoning abilities and thought processes utilized by a human fact verifier to find evidence relevant to natural numerical claims.

We create an effective and realistic fact verification dataset for numerical claims by using the QuanTemp[65] dataset as a foundation, as it contains substantial natural numerical claims and their corresponding fact verification labels. A straightforward method to achieving such a dataset would be through crowdsourcing; however, obtaining relevance labels for a significant number of claims is expensive and requires expert annotators. Hence we resort to weak supervision. We develop a three-step pipeline: an Automatic Query Generator, followed by a Search Engine Results Page Processor (SERPP), and finally, a web crawler to curate our dataset. The full data collection pipeline with an example is shown in Figure.3.1.

The subsequent subsections will first detail each of the components of our pipeline, following which we describe our approach to evaluate the dataset's quality.

Figure 3.2: Example of query generation output from LLM with redundant and irrelevant items.

3.1.1 Automatic Query Generation

Fact verifiers often search the web or private documents to fetch relevant evidence articles to verify claims. To emulate their thought process, we use few-shot learning to generate these web search queries using the justification document provided for each claim. We hypothesize that utilizing a claim's justification document in a few-shot learning setup would generate both explicit and implicit queries required to verify a natural numerical claim. The prompt to generate Web search queries includes the claim, its publication date, justification document, and one static example crafted manually by the authors. The instruction specifies the generator to produce queries that sufficiently address the numerical aspects of the claim. The full prompt can be found in the Appendix.A.2.

A widely known problem of LLMs is hallucination. Additionally, LLMs also tend to produce redundant information within the response. An example of this is shown in Fig.3.2. To mitigate these issues, we put in place a filtering process after query generation. This step ensures we only keep queries that are both unique and pertinent. The filter is constructed using the relevance scores provided by the Maximal Marginal Relevance (MMR) algorithm. Given that the queries generated may be explicit or implicit we need to ensure that the selected queries are still relevant but don't deviate from the claim's core essence. Therefore, we employ a weighted average approach, to combine the scores with respect to the claim and justification document to form the end scores for each query. Finally, we apply a threshold to retain only the most relevant and independent queries for each claim. Hence, for a given claim C, justification document J and their corresponding sub-queries $q_1, \ldots, q_k$, the final set of queries Q is formed by:

$$s_{q_i} = \alpha \cdot \text{MMR}(q_i, C, \lambda) + (1 - \alpha) \cdot \text{MMR}(q_i, J, \lambda)$$

$$Q = \{q_i \mid s_{q_i} > \gamma \text{ for } i \in \{1, k\}\}$$

Here, α is the coefficient used to control the importance needed to be given to relevance with respect to the claim versus that to the justification document, λ is the diversification factor used in MMR and γ is the threshold for selecting a query.

3.1.2 SERPP

We make the assumption that evidence for claims can be fetched and accessed over the public web. We only fetch web search results that were published before the claim was made. Each result contains the URL to the full web page, a snippet containing the most relevant phrases in the web page and the publication date of the web page. The results of the web search can be irrelevant to the query if there are no significant results found for it (See Figure.3.3). To exclude such instances in creation of the relevant set of results for a claim, we filter the results by measuring the similarity between the result snippet and the corresponding query used to perform the Google search. Then, we utilize a language model pre-trained for text relevance to encode both the query and text snippet and take the cosine similarity between them to form the final relevance score. Once we obtain the relevance score, we set a threshold to filter out noisy results from being included in the set of relevant records for a claim.

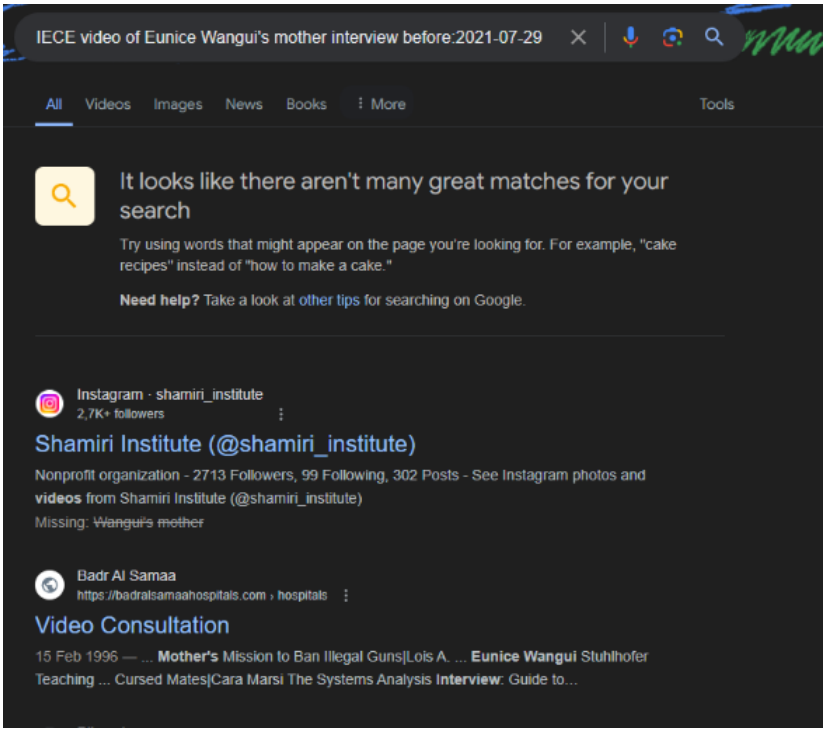


Figure 3.3: Example of query generation output from LLM with redundant and irrelevant items.

3.1.3 Crawling

Many works that use web search results as evidence use the web search snippets directly as evidence for the corpus[5, 65]. However, these snippets lack context or important information such as disclaimers that may be useful for the fact verification component. For example, several claims originate from satirical web pages that specifically add disclaimers within their page. While we also use these snippets in our work, we crawl all the web pages from our web search

results to enrich our corpus for future work. We then parse the HTML results to store the main text content. We exclude any links that are unavailable, behind paywalls, or forbidden.

3.1.4 Distractors and Leakage

The mapping of relevant records for each claim, resulting from the above components, forms the query relevance judgments, or *qrels*, of our dataset. A realistic corpus contains sufficient distractors that are close to the set of relevant records but are not relevant. Hence, the final corpus includes all the relevant and noisy filtered records resulting from the previous components, while the *qrels* only maps claims to their relevant record. To further increase the difficulty of the dataset, we feed the claim verbatim to SERPP with the temporal filter as mentioned before, and add the top-10 records and their corresponding crawled web pages to the corpus. To prevent gold evidence leakage, we evict from the corpus and *qrels*, all the records whose links originate from fact-checking websites. We use the set of fact-checking domains provided by QuanTemp for the same.

Split	**TRUE**	**FALSE**	**CONFLICTING**	**Total**
Train	1824	5770	2341	9935
Validation	617	1759	672	3084
Test	474	1423	598	2495
			Total	15514

Table 3.1: Distribution of claims by veracity label

Split	**Temporal**	**Interval**	**Statistical**	**Comparison**
Train	2672	1541	4660	1051
Validation	840	469	1432	339
Test	681	347	1210	255
Total	4193	2357	7302	1645
%	27.06	15.21	47.12	10.61

Table 3.2: Distribution of claims by numerical abilities required

3.1.5 Dataset Statistics

The Quantemp dataset and hence our dataset consists of a total of 15, 514 natural numerical claims. The dataset is unbalanced as a majority of claims are labeled as *False*. The distribution of the dataset by veracity label is provided in Table.3.1. The numerical claims in the dataset are further divided by the numerical abilities required to verify them namely- temporal, statistical, interval, and comparison. Majority of the claims require statistical computational abilities followed by claims that required temporal understanding (See Table.3.2). Our evidence corpus consists of 165.7k records. The average snippet length is about 154 tokens and the average number of snippets relevant to each claim is about 6.35.

The Automatic query generator outputs a variable number of queries required to retrieve relevant evidence per claim. The average number of queries per claim is 6.76, which is much higher than other methods such as PgmFC[43] and ClaimDecomp[9]. Additionally, the average number of tokens per query is 8.51 which is much lower than those present in the queries generated by PgmFC[43] and ClaimDecomp[9].

3.2 Evaluation of Query Planning Methods

QUANTEMP++ serves as an ideal dataset to evaluate the retrieval performance of various query planning methods in the task of verification of natural numerical claims as it contains the relevance labels produced by weak supervision and offers an open domain setting for retrieval. Additionally, the corpus contains evidence that addresses both implicit and explicit aspects of a natural numerical claim. Since the corpus does not contain gold or temporal leakage, we can make a realistic evaluation of key query planning methods with this dataset.

3.2.1 Key Query Planning Methods

In section 2.2.1, we identified three key approaches for querying evidence to address the different aspects of a natural claim. The first strategy used the claim directly to query for evidence documents. The second and third strategies decomposed the claim into sub-queries and then used these sub-queries to retrieve evidence. The second strategy decomposed the claim to sub-queries with a smaller fine-tuned model while the third used prompting strategies to generate these decompositions. Decomposition has previously shown to improve recall by explicitly addressing the different aspects of a natural claim[9, 43, 65]. However, the technique used to decompose a claim into sub-queries can have varying downstream retrieval impacts. Hence, we evaluate the retrieval performance of these three paradigms using the QUANTEMP++ dataset to observe their strengths and weaknesses in addressing the different aspects of natural numerical claims.

3.2.2 Evidence Aggregation for Decomposed Queries

In the above section, we introduce three query planning methods, in two of which utilize decomposition. The latter two techniques break down a claim into its sub-queries and retrieve evidence for each sub-query. Once the documents are retrieved, a systematic approach to aggregate these results to form the final evidence set is required. In Section.2.2.1 we saw that previous research in information fusion demonstrated the efficacy of the CombMAX and CombSUM fusion functions that could potentially overcome the above challenge. We evaluate and compare the above methods using the oracle queries, corpus, and relevance labels from QUANTEMP++ using a zero-shot retriever. For a given claim with sub-queries $q_{1..K}$ and a corpus $\mathcal{D}$ let r_{ki} be the score assigned by the retriever for the sub-query q_k for document d_i in the corpus. Hence, the 6 settings we evaluate are:

- TOP-1: Forms the final set by taking the top-scored document for each sub-query.

$$\text{evidence\_set} = \left\{ d_i \mid d_i = \arg\max_{d \in \mathcal{D}} r_{kd} \right\}_{k=1}^{K}$$

- COMBMAX: Using this strategy, each document in the corpus is assigned and ranked by the maximum score achieved by any of the sub-queries.

$$\text{COMBMAX}(d_i) = \max_{j=1,\ldots,k} r_{ji}$$

- COMBMAX-NORM: Using this strategy, the scores assigned by the retriever to the documents are first normalized per sub-query and then the COMBMAX function is applied to them to get their final score.

$$r'_{ki} = \frac{r_{ki} - \min_{d\in\mathcal{D}} r_{ki}}{\max_{d\in\mathcal{D}} r_{ki} - \min_{d\in\mathcal{D}} r_{ki}} \tag{1}$$

$$\text{COMBMAX-NORM}(d) = \max_{i=1,\ldots,k} r'_{di}$$

- COMBSUM: Using this strategy, each document in the corpus is assigned and ranked by the sum of scores achieved by any of the sub-queries.

$$\text{COMBSUM}(d_i) = \sum_{j=1,\ldots,k} r_{ji} \tag{2}$$

- COMBSUM-NORM: This strategy first normalizes the document scores by each sub-query as in (1) and then assigns each document the sum of scores across sub-query as the final score as with (2).
- CONCAT: We take the concatenation of the sub-queries for the final query for our retrieval.

3.2.3 Retrieval Performance of query planning methods

We analyze the results of the above evaluation of the aggregation method and select the one that performs the best using a zero-shot retriever. Now using this we evaluate our key query planning methods discussed in 3.2.1 using the claims, corpus, and relevance labels of QUANTEMP++. Since the corpus contains various distractors, we also assess the efficacy of adding external signals to reduce noise by applying a temporal filter to the retriever. Hence, when a query for a claim is made, the retrieval for each of these queries is restricted to documents published before the claim.

3.2.4 Downstream Impact

Observing the corresponding downstream impact of each of the query planning methods is crucial as it represents the final outcome of the pipeline. Hence, for each of the query planning methods, we first retrieve evidence, aggregate, and use the top evidences to train and test the downstream NLI classifier. A Multi-Genre Natural Language Inference (MNLI) model is used to train and predict the veracity label of the claim as True, False, or Conflicting. Since most MNLI models have a limited context length, we limit the input length of the evidence to the top-k fetched snippets. Hence, given the claim and the top-k retrieved snippets by the corresponding query planning method we train and test this model to predict the veracity label of the claim.

Chapter 4

Experiments

In this section, we detail the implementation of our methodologies and the evaluations we carry out in our work. We first explain the implementation of the data creation pipeline of QUANTEMP++ and how we evaluate its quality. Next, we outline our experimental setup to assess the retrieval and downstream performance of key query planning strategies for verifying natural numerical claims.

4.1 Data Creation

To create QUANTEMP++ we utilized a three-component pipeline that utilizes the claims and their corresponding justification document to form a corpus of evidence snippets and relevance labels. We source the claims and their veracity labels, justification documents, and metadata from QuanTemp[65]. The three components of the data creation pipeline are the Automatic Query Generator(AGQ), the Search Engine Results Page processor (SERPP), and the Crawler. The execution details of each of these phases are as follows:

Automatic Query Generator

To generate the queries that address the different implicit and explicit information needs of a numerical claim, we first prompt an LLM and then use a filter to enforce diversity and relevance. Open AI's GPT models have shown to be effective in simulating user search queries[34, 51], hence we use the GPT-3.5-TURBO model to generate the questions. The filter utilizes the MMR algorithm to enforce diversity and then measure relevance to the claim and its corresponding justification document. Within the MMR algorithm, a language model is used to score text similarity between documents, as well as between a document and a claim and its justification documents. We employ the *paraphrase-MiniLM-L6-v2*[48] model to assess the similarity between texts. After tuning, we finally set the diversification constant λ of the MMR algorithm to 0.4, the weighted average coefficient α to 0.8, and the filter threshold γ to 0.4.

4.1.1 Search Engine Page Processor (SERPP)

Once we have the queries generated per claim, we use these queries to search the web for evidence items and filter these items based on their relevance to the query as a post-processing measure. We use SerpAPI[1] to carry this out. Given a query, SerpAPI[1] fetches the top 10 web pages on the Google Search Results page. Each result consists of a URL to the web page, a search result snippet, and a publication date. To prevent temporal leakage, we add the *before:* filter to the query, which restricts the results to those published before the date provided. Additionally, the rankings are also reverted to the state on the day provided. To filter irrelevant results, we use *all-mpnet-base-v2*[61] which is a pre-trained general-purpose model used by several works to calculate relevance[19, 6]. We set the threshold of the filter to 0.45. The relevance labels are formed by mapping these filtered results of every sub-query to the claim whereas the final corpus includes all unfiltered results. Distractors are also added to the corpus in a similar way; however, instead of sub-queries, the full claim is provided with a temporal filter but without any post-processing filtering.

4.1.2 Crawler

Given the filtered set of search results for each sub-query of the claim, we crawl the URLs of these results on a best-effort basis for future researchers to use. We utilize the *grequests* library to perform concurrent queries on the links and retrieve the HTML content. We limit the response to text/HTML or XML types and restrict the language to English using request headers. We parse the resulting HTML and extract the main content using the article extractor of the *boilerpy* library. We exclude any results that respond with error codes due to authorization, client, or server-side errors.

4.1.3 Evaluation

To evaluate the quality of our dataset, we conduct manual validation to verify the main hypotheses used to generate our dataset. Additionally, we perform quantitative analysis by measuring the overall impact of our dataset on the downstream task of fact verification of numerical claims.

Manual Validation

The generation of our dataset involves two main hypotheses:

1. We hypothesize that utilizing the justification document along with the claim to generate Google search queries through few-shot learning produces high-quality queries mimicking the thought process of human fact verifiers.
2. Additionally, we hypothesize that these queries generated help produce high-quality evidence, both implicit and explicit when used to search the open web.

We evaluate the above hypotheses through manual validation. We select 50 samples from our dataset randomly and present their artifacts to annotators. These samples are selected to reflect the original distribution of the dataset along the dimensions of claim complexity and veracity label. To assess the first hypothesis, we present the annotator with the claim and the final generated and filtered queries that are output from the AGQ component in our pipeline. We

additionally provide the justification document for each sample for reference. For each of these generated queries per claim, the annotator needs to mark it as relevant or irrelevant in terms of its aptness for verifying the claim. Additionally, for each set of queries per claim, the annotators score the set as a whole in terms of comprehensiveness and redundancy on a 5-point numerical scale. We adopt the following definitions from ClaimDecomp[9]:

- **Redundancy**: The query set should be as minimal as is practical and not contain repeated queries. Hence, a high score of 5 would mean that the set of queries is not concise and only repeatedly focuses on a few aspects required to verify the claim.
- **Comprehensiveness**: Queries are defined to be comprehensive if they cover as many aspects of the claim as possible. A high score of 5 would mean the given set of queries covers all aspects, both implicit and explicit required to verify the claim.

In our task instruction for annotators, we include a hand-crafted example as examples have proven previously to improve the annotator's understanding of the task[40]. The instructions also include explanations for each numerical rating, detailing what each score signifies regarding comprehensiveness and redundancy. We have three annotators, two familiar with the project, and one unfamiliar to annotate the samples.

For our second hypothesis, we follow the same setting as above but present the annotator with the claim and relevant Google search snippets generated and filtered that are output from the SERPP component. Similar to annotating the queries, the annotators now mark each snippet as relevant or irrelevant along with scoring each set of snippets as a whole in terms of comprehensiveness and redundancy. While we do have the final corpus of full-length web documents, we refrain from evaluating them as fetching granular relevance labels for web documents has proven to be very challenging in previous studies and has resulted in low annotator agreements[71, 28]. An example annotation is shown in Fig.4.1 and the full task instruction can be found in Appendix.A.3.

Task 2

Claim	Published	Just Doc	Summary	Snippet	Relevant	Completeness	Redundancy
"23,344 mail-in ballots came from people who no longer lived at that address. … … 284,412 ballot images were, quote, corrupt; they quoted 'corrupt or missing.' Oh, but I only lost by a little more than 10,000 votes."	2021-10-12	Comment on thi	In Arizona, the C	AP report: Few AZ voter fraud cases, discrediting Trump's claims - PBSPHOENIX (AP	☑		
				Donald Trump Perry, Georgia Rally Speech Transcript September 25284,412 ballots	☑		
				Trump Loses Arizona—Again - WSJFormer President Trump claims Arizona's ballot a	☑		
				PA Lawmakers: Numbers Don't Add Up, Certification of Presidential …Among the 6,	☑		
				Michigan Presidential Election Voting History - 270toWin.comInformation on how t	☐		
				After 2020 losses US Republicans move to limit voting rightsFormer President Dona	☑	5	1

Task 1

Claim	Pubblished	Justification Dc	Summary	Querries	Relevant	Completness	Redundancy
"23,344 mail-in ballots came from people who no longe	2021-10-12	Comment on thi	In Arizona, the C	How many mail-in ballots were reported to have come from people who no longer lived at their address in Arizona?	☑		
				What was the reported number of "corrupt or missing" ballot images in Arizona according to Trump's claims?	☑		
				How many votes did Trump claim to have lost by in Arizona?	☑		
				How many more votes than voters did Trump claim were reported in Philadelphia, Pennsylvania?	☑		
				How many people were listed on the voter rolls in Michigan, according to Trump, who hadn't voted in over 20 years?	☑		
				In which states did Trump claim to have won while losing the election, according to the passage?	☐	5	1

Figure 4.1: Example of manual annotation conducted to assess the quality of QUANTEMP++

Metrics: We aim to evaluate the set of queries and Google search snippets generated for comprehensiveness and redundancy. We average the scores of comprehensiveness and redundancy across annotators to get a final score per claim. We then check the frequency distribution of these scores to draw analysis independently for each aspect and each task. We also aim to quantify the amount of noise being generated in our pipeline. Hence, we calculate the precision

of relevance labels of queries per claim. We then take the average of the precision score across claims to get the final precision. We carry out the same procedure for quantifying the noise in evidence snippets. To assess the reliability of the annotations, we calculated Fleiss' kappa[17] for each of the annotation tasks.

4.1.4 Quantitative Analysis

To evaluate the impact of our dataset in the downstream task of fact verification of natural numerical claims, we fine-tune a veracity prediction model using the train and validation splits of our dataset. We then evaluate the performance of this model against the test split of our dataset to measure the gain in knowledge our dataset could have imparted to the veracity classifier. We compare the performance of the following models:

1. NAIVE-CLASSIFIER: A Veracity Classifier that only predicts the majority class of the dataset.
2. ROBERTA-MNLI-CLAIM-NO-EV: A pre-trained MNLI model fine-tuned to predict the veracity of the claim only by having just the content of the claim as input.
3. GPT-3.5-TURBO: A few-shot learning model that given the claim and oracle evidence(search snippets), predicts the veracity of the claim. The few-shot learning model dynamically selects one example per veracity class from the training set, similar to V et al. [65]. We use the *gpt-3.5.-turbo* generative model for this purpose.
4. ROBERTA-MNLI-QTEMP++: A pre-trained MNLI model fine-tuned to predict the veracity label of the claim given the claim and its corresponding oracle evidence snippets.
5. ROBERTA-MNLI-GOLD: A pre-trained MNLI model fine-tuned to predict the veracity label of the claim given the claim and its justification document. This model serves as the upper bound that can be achieved with the pre-trained MNLI model for QUANTEMP++.
6. GPT-3.5-TURBO-GOLD: A zero-shot model that predicts the veracity of the claim given the claim and its gold justification document provided by a human fact-checker. This model serves as the upper bound that can be achieved with the *gpt-3.5.-turbo* model for QUANTEMP++.

MNLI Classifier

For the MNLI classifier, we follow [65] and utilize the *roberta-large-mnli* model. This classifier consists of an input encoder initialized with the *roberta-large-mnli* model, followed by a multi-class classification head. The maximum input length of this model is 512 but we set the limit to 256 due to resource constraints. A summary of the model can be found in Appendix.A.4. We fine-tune all the MNLI models using the Adam optimizer with a weight decay of $1 \times e^{-5}$ and a learning rate of $2 \times e^{-5}$. The training process utilizes the cross-entropy loss function and incorporates early stopping with a patience of 2.

Metrics

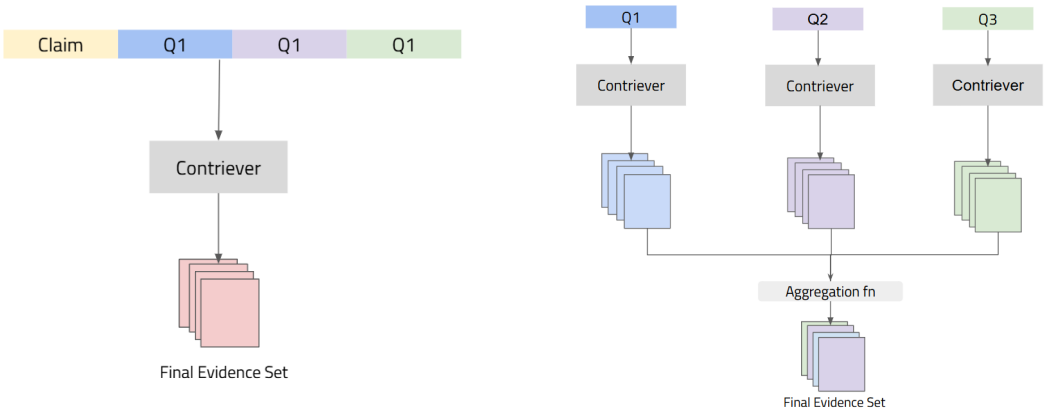
For each of these settings, we evaluate the accuracy, per-class F1, macro-F1 (M-F1), and weighted-F1 (W-F1) scores to account for the class imbalance in the dataset.

4.2 Evaluation of Query Planning Methods

In Section.3.2.1, we introduced three key query planning methods to retrieve evidence relevant to verifying numerical claims. We evaluate the retrieval performance of these three key query planning methods using the following settings:

- **CLAIM-ONLY**: In this setting, we use the entire numerical claim to retrieve evidence from the corpus.
- **CLAIMDECOMP**: ClaimDecomp[9] utilized crowd workers to generate training data to decompose a claim into yes or no questions. This dataset serves rich examples for claim decomposition where both implicit and explicit aspects are addressed. Therefore, we utilize this dataset to prompt GPT-3.5-TURBO with dynamically selected in-context samples. We set the temperature of the LLM to 0.1 as we need relatively deterministic answers for our decomposition task, ensuring our outputs are restricted to the concepts in our input.
- **PGMFC**: PgmFC[43] generates step-by-step instructions derived from the decomposition of the original claim which can be beneficial to fetch evidence to conduct numerical reasoning about the claim. Hence, we incorporate the prompting strategy used by the original paper to generate sub-queries[43].
- **QGEN**: QUANTEMP++ consists of oracle queries generated by prompting an LLM with the claim and its justification document. Although these decompositions are not human-annotated, the claim-subquery pairs contain the thought process utilized by a human fact checker. We attempt to distill this knowledge into a smaller model by using these pairs as training examples. We use Google's FLAN-T5-LARGE[38] instruction-tuned model for this purpose. We train the model with early stopping with a patience of 2, a learning rate of $2 \times e^{-5}$, and a batch size of 8.

To form an upper bound on the performance of decomposition-based methods we introduce the ORACLE-QUERIES setting that uses the oracle queries from QUANTEMP++ to retrieve evidence.



(a) CONCAT (b) Other aggregation methods

Figure 4.2: Formation of the final evidence set using the (a) CONCAT strategy and (b) the other 5 strategies like TOP-1, COMBMAX and COMBSUM.

Evaluation of Aggregation methods

For the decomposition-based query planning methods of CLAIMDECOMP, PGMFC and QGEN we saw there was no principled study to aggregate the results in a principled manner, however, we saw 6 result aggregation strategies namely TOP-1, CONCAT, COMBMAX, COMBMAX-NORM, COMBSUM and COMBSUM-NORM. Since we cannot evaluate every combination exhaustively, we first evaluated the result aggregation strategies using the oracle queries in QUANTEMP++ using a zero-shot retriever. For CONCAT, we concatenate the claim and queries with a separator and use this as input to our retriever to obtain the documents that form the final evidence set. For TOP-1 we selected the top-scored evidence document per query to form the final evidence set. For other strategies, for each claim, we first retrieved documents per query and then applied the aggregation filter to form the final set of evidence documents (See Fig. 4.2). We then evaluated the performance of this evidence set against the relevance labels in QUANTEMP++. For the zero-shot retriever, we used Contriver[25], a BERT-based model pre-trained for information retrieval. CONTRIEVER is a dual encoder model pre-trained with contrastive learning that performs well across different topics. We utilize FAISS[29] to create our corpus index, employing cosine similarity to retrieve related documents and their respective scores for each query. We evaluate NDCG@10 and Recall@100 to understand the retrieval performance of these aggregation methods.

Evaluation of the Retrieval

Once, we analyze the performance of the aggregation methods, we select the best-performing method and then evaluate the above-mentioned query planning methods to gain insights. We again use the same zero shot retrieval setting with CONTRIEVER[25] and FAISS to carry out our experiments. We additionally add a temporal filter in a second setting where all documents with dates larger than that of the claim are filtered out after retrieving documents from the index. We assess NDCG@k, Recall@k, Mean Reciprocal Rank (MRR), and Precision@k to evaluate the query planning's effectiveness in ranking documents, prioritizing relevant items, covering diverse aspects, and handling noise at the retrieval of documents for numerical claims.

Evaluating Downstream Impact

To evaluate the downstream impact of the different query planning methods, we take the final set of evidence snippets aggregated from the retrieval stage and use the top 3 snippets with document scores above 0.5, and the claim as the input to the MNLI classifier. We use this setting to both test and train the classifier. For the classifier, we use the same *roberta-large-mnli* model as mentioned in Section.4.1.4, as well as the same fine-tuning settings. We observe the metrics of accuracy, Macro, and Weighted F1 to gauge performance across the query planning methods. We additionally use the numerical taxonomy of the natural claims (3.2) in QUANTEMP++ to analyze downstream impact per class.

4.3 Ablations

To evaluate the aggregation techniques and mainly the performance of different query planning methods, we chose CONTRIEVER[25] due to its superior performance in retrieval in previous works and its efficacy at retrieving relevant evidence for queries from diverse domains. This aspect is beneficial in our case as our dataset consists of numerical claims from different domains like politics, medicine, science, etc. Regardless, we test the efficiacy of a few other retrieves, namely BM25[50], ANCE[80], DPR[31], and Tas-b[23]. There are several other retrievers utilized in different automatic fact verification pipelines but due to time constraints, we only include the above mentioned ones in our studies.

Post our experiments to evaluate the retrieval performance of key query planning methods, we assess its downstream impact by feeding the top-k retrieved evidence to the NLI model. To understand the best k value, we finetune our MNLI model with different number oracle evidence snippets per claim. Specifically, we experiment with k values of 1, 3, 5, 7, and 10. We abstain from going above 10 as we set the maximum input limit of *roberta-large-mnli* as 256 tokens. We additionally observe the retrieval performance for ORACLE-QUERIES for the above-extended k values for the validation set.

Hardware configuration

The experiments were conducted on a dedicated private server running on Arch Linux. This platform features a 16-core 2nd Gen AMD EPYC™ 7302 processor paired with two NVIDIA GeForce RTX 3090 GPUs, offering significant computational capabilities. Additionally, the server is equipped with 256 GB of RAM, ensuring smooth and uninterrupted performance during intensive computational tasks.

Chapter 5

Results and Discussion

In this section, we present the results of our experiments and discuss the insights gained from them. First, we assess the quality of the dataset created–QUANTEMP++, and then we utilize this dataset to evaluate the retrieval performance of key query planning methods. Finally, we discuss the downstream impact of these query planning methods. Through the insights gained from the above, we intend to answer our three research questions:

- **RQ1**: Does query decomposition help retrieve quality evidence from the web for the verification of natural numerical claims?
- **RQ2**: How do existing query planning methods perform in terms of retrieval of relevant evidence snippets to verify numerical claims?
- **RQ3**: What is the downstream impact of these query planning methods on the task of verification of numerical claims?

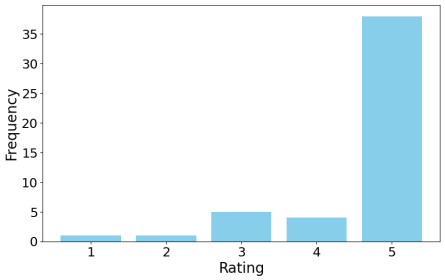
5.1 Dataset Quality

To answer **RQ1**, we evaluate the quality of the QUANTEMP++ dataset both qualitatively and quantitatively.

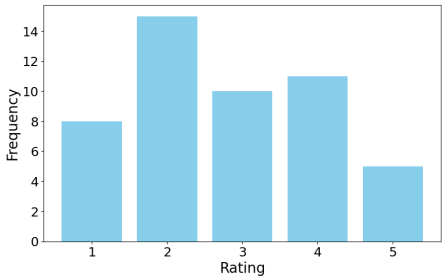
5.1.1 Qualitative Analysis

To evaluate the two main assumptions in the design of our dataset creation pipeline, 3 researchers manually annotated 50 samples from the dataset. Across both tasks of annotating queries and snippets for completeness and relevance we get significant agreement (See Table.5.1). However, due to the more subjective nature of redundancy, our agreement for both tasks in this aspect is only moderate. From these annotations, we gain the following two main insights from this evaluation:

Firstly, we see that utilizing a few-shot learning setup effectively generates queries to verify natural numerical claims. The frequency distribution of numerical scores assigned by our annotators for the set of queries generated per claim is provided in Table.5.1. The results of the manual evaluation show that generated queries cover most of the implicit and explicit numerical

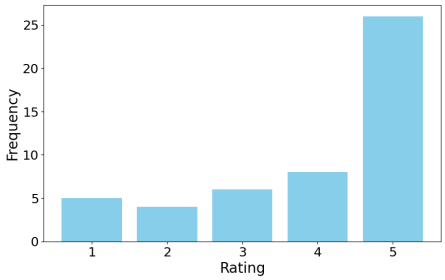


(a) Comprehensiveness

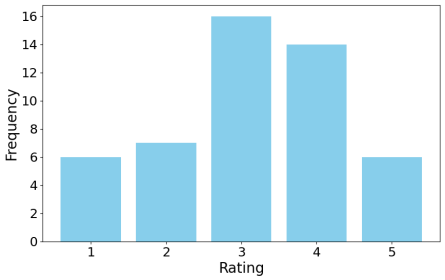


(b) Redundancy

Numerical scores assigned for the set of queries generated.



(c) Comprehensiveness



(d) Redundancy

Numerical scores assigned for the set of evidence snippets.

Figure 5.1: Distribution of numerical scored assigned by annotators during qualitative analysis of QUANTEMP++.

aspects of the natural claim 86% of the time. However, these queries exhibit a significant amount of redundancy, with approximately 84% of the cases containing some redundancy. We also see that the average precision of these queries is 90%, meaning most of our queries are relevant to the claim. It is also to note that the 10% of queries that are irrelevant could have potentially impacted downstream quality negatively in the data creation pipeline.

Metric	Queries	Evidence Snippets
Completeness	0.61	0.66
Redundancy	0.4	0.45
Relevance	0.58	0.67

Table 5.1: Inter-annotator agreement scores for the tasks in the qualitative analysis of QUANTEMP++

Secondly, using the above-generated queries to search the web generally produces evidence snippets to verify natural numeric claims. The frequency distribution of numerical scores assigned by our annotators for the set of evidence snippets generated per claim is provided in

Table.5.1. Manual evaluations reveal that the mapped evidence snippets mostly cover different aspects of the claim 69% of the time, and partially cover them an additional 12% of the time. However, for 87% of cases, the snippets contain redundant information. We also see that the average precision of the evidence snippets is 62%. A manual analysis of 10 claims with low coverage scores reveals that this issue arises because parts of the necessary evidence are not available on the public web or are based on multimodal information.

5.1.2 Quantitative Analysis

To quantitatively measure the efficacy of our dataset, we fine-tune ROBERTA-LARGE-MNLI with the training split of our dataset and evaluated its performance on our test split of oracle claim and evidence pairs. From Table.5.2 we see that the model trained on our dataset, ROBERTA-MNLI-QTEMP++, performs much better than all the baselines on all metrics. It achieves relative gains of about 41% over NAIVE-CLASSIFIER and about 16% over the ROBERTA-MNLI-CLAIM-NO-EV classifier. We also see that the performance of ROBERTA-MNLI-QTEMP++ is closer to ROBERTA-MNLI-GOLD indicating that we are very close to the upper bound that we can achieve using the ROBERTA-LARGE-MNLI model.

NLI Classifier	**Accuracy**	**Per class F1**			**Total**	
		T	F	C	W-F1	M-F1
NAIVE-CLASSIFIER	57.03	0.00	72.64	0.00	24.21	41.42
GPT-3.5-TURBO	54.15	20.88	37.81	20.88	52.87	43.44
GPT-3.5-TURBO-GOLD	62.32	56.67	75.35	28.00	60.47	53.37
ROBERTA-MNLI-CLAIM-NO-EV	63.2	24.71	79.93	47.37	61.53	50.67
ROBERTA-MNLI-QTEMP++	**65.25**	**79.92**	**47.76**	**49.67**	**66.56**	**59.11**
ROBERTA-MNLI-GOLD	69.66	56.86	82.92	48.79	69.79	62.85

Table 5.2: Performance of different claim veracity prediction models on QUANTEMP++.

Surprisingly, our smaller ROBERTA-MNLI-QTEMP++ classifier achieved relative gains of approximately 36% over GPT-3.5-TURBO, despite the latter having 520 times more parameters. Such phenomena have been observed in recent work that smaller language models specifically finetuned for certain NLP tasks provide comparable or better performance than larger GPT models [35, 76]. Brown et al. [8] also suggests that one reason GPT-3.5-TURBO might struggle with NLI tasks is due to its autoregressive modeling design, which may be incompatible with the requirements of NLI tasks. Additionally, it is to be noted that the performance of GPT-3.5-TURBO-GOLD falls significantly short of that achieved by ROBERTA-MNLI-QTEMP++, hinting that in a few shot learning setting GPT-3.5-TURBO is unable to comprehend the complex numerical information in the input to provide accurate results. Conversely, ROBERTA-MNLI-QTEMP++, despite being a smaller model, due to fine-tuning, excels at understanding the statistical patterns in the input, allowing it to internally reason about them and provide accurate predictions.

5.1.3 Key takeaways

In summary, the final takeaways from our qualitative and quantitative analysis of QUANTEMP++ are as follows:

1. Utilizing a few-shot learning setup effectively generates queries to verify natural numerical claims.
2. Using the above-generated queries to search the web generally produces evidence snippets to verify natural numeric claims.
3. Smaller models such as ROBERTA-MNLI-QTEMP++ when trained for specific tasks, especially those involving complex reasoning are capable of outperforming other baselines, including the general-purpose GPT-3.5-TURBO, which is 520 times larger.

5.2 Impact of Query Planning Methods

To answer **RQ2**, we first evaluate various aggregation techniques using the oracle queries in QUANTEMP++ and then use then utilize the best aggregation method to evaluate retrieval for key query planning methods. The below sections delineate the results and insights of these evaluations

5.2.1 Aggregation of retrieval results from decomposed queries

In order to compare the retrieval performance of differenct claim decomposition techniques, we require a principled method to aggregate the retrieval results across the decomposed queries for each method. The necessity for a principled aggregation technique becomes further pronounced when different decomposition methods produce a variable number of queries per claim. Table.5.3 shows the retrieval performance of 6 aggregation techniques against the QUANTEMP++ dataset with oracle queries.

Method	NDCG@10	Recall@100
TOP-1	0.42	0.37
CONCAT	0.47	0.67
COMBSUM	0.45	0.75
COMBMAX	0.50	0.79
COMBSUM-NORM	0.49	0.80
COMBMAX-NORM	**0.57**	**0.82**

Table 5.3: Retrieval performance of different aggregation techniques on QUANTEMP++

From the results, we see that COMBMAX-NORM performs the best whereas TOP-1 performs the worst with respect to the NDCG and Recall. The limitation of the TOP-1 aggregation technique arises from the reliance on the retriever to give the perfect result on the topmost position for each decomposed query of the claim, exaggerating any errors in decomposition. Each

decomposed query, especially when underspecific, can miss out on addressing important aspects of the claim by only considering the first result.

COMBSUM-NORM does better than TOP-1 but significantly worse than COMBMAX-NORM. In previous works COMBSUM is incorporated in information fusion tasks in order to take advantage of the Chorus Effect [70]. While the Chorus Effect of this technique works to push evidence relevant to all queries on top to reduce outliers, it also actively discourages diversity. Additionally, if the decomposed queries are redundant, this effect will further suppress diversity. The presence of queries addressing implicit aspects may not be related to one other causing confusion in the aggregation process. On the contrary, COMBMAX operates under the premise that the system of decomposed queries each excels at addressing their own aspects, a phenomenon known as the Dark Horse Effect [70]. Since this phenomenon promotes diversity in retrieval, which is important for the verification of claims, especially numerical, we see significantly higher performance in NDCG and recall.

Lastly, we see normalization on input retrieval scores per query during aggregation improves the final retrieval performance. As mentioned by Wu et al. [78], normalizing input scores allows equal opportunity for each decomposed query to contribute to the results regardless of the variation in their scoring of evidence.

5.2.2 Retrieval Results of Key Query Planning Methods

We evaluated the retrieval performance of the various query planning methods on our dataset. The results are listed in Table.5.4. From these results, we obtain the following main insights.

Query mode	NDCG@10	Recall@10	Recall@100	P@10	MRR
CLAIM-ONLY	0.31	0.30	0.51	0.19	0.50
Decomposition					
ORACLE-QUERIES	0.54	0.51	0.82	0.32	0.651
CLAIMDECOMP	**0.31**	**0.30**	**0.56**	**0.19**	**0.492**
PGMFC	0.30	0.29	0.52	0.18	0.481
QGEN	0.29	0.27	0.54	0.16	0.481

Table 5.4: Retrieval performance of different query planning methods using COMBMAX-NORM on QUANTEMP++

Different query planning methods have varying impacts on the final retrieval results which can be seen in Table.5.4. From these results, we observe that although the NDCG@10 scores across all query planning methods show minimal variation, Recall@100 for decomposition-based query planning, particularly for CLAIMDECOMP, exhibits a notable increase with relative gains as high as 10% when compared to CLAIM-ONLY. Additionally, Recall@100 for PGMFC and QGEN shows increases of 2.4% and 5.9% respectively. Previous research has demonstrated that using only the claim to retrieve evidence exhibits low recall and hence coverage of aspects required to verify the claim. Decomposition methods that generate queries to address various aspects indeed improve coverage. Ensuring evidence coverage for a claim is crucial in the initial

stage of retrieval, particularly for numerical claims. Missed aspects in this stage can often lead to completely different NLI results. An example of this is given in Appendix.A.1.

In contrast to Recall@100, the Recall@10 for PGMFC and QGEN are slightly worse than CLAIM-ONLY, while Recall@10 for CLAIMDECOMP is comparable to CLAIM-ONLY. Additionally, Precision@10 for PGMFC is minimally lower than CLAIM-ONLY by 0.8%, and QGEN is lower by 2.4%. This could be attributed to noisy results being included from the decompose, retrieve, and aggregate process for these methods. To reduce noise, we added a simple temporal filter to restrict retrieval of evidence to ones that were only published before the claim. This technique improved the results of all query planning methods slightly (See Table.5.5). However, we see that the gap in performance of CLAIM-ONLY versus decomposition methods widened for Recall@100. Adding this filter helped improve Precision@10, especially for that of QGEN by about 3%. Additionally, the MRR for QGEN increased by 9%. Since most downstream Natural Language Inference (NLI) tasks are sensitive to the order of results, especially during fine-tuning[60], this increase in MRR could be crucial. Overall we see that adding external signals like temporal relevance to the retrieval pipelines can potentially reduce noise and improve the overall performance of query planning methods, especially those based on decomposition.

Query mode	**NDCG@10**	**Recall@10**	**Recall@100**	**P@10**	**MRR**
CLAIM-ONLY	0.33	0.32	0.51	0.20	0.54
Decomposition					
ORACLE-QUERIES	0.57	0.54	0.82	0.34	0.68
CLAIMDECOMP	0.34	0.32	0.57	0.21	0.53
PGMFC	0.32	0.31	0.52	0.20	0.52
QGEN	0.33	0.31	0.56	0.20	**0.57**

Table 5.5: Retrieval performance of different query planning methods using COMBMAX-NORM on QUANTEMP++ with temporal filtering.

As a part of future work, a more balanced aggregation method and incorporation of background knowledge in the decomposition phase may be needed to catch up to the oracle setting. While utilizing decomposition-based query planning, given the retriever is frozen, noise could be introduced by faulty decomposition and/or by using COMBMAX-NORM to aggregate results. It is to be noted that all the decomposition-based methods of CLAIMDECOMP, PGMFC, and QGEN rely solely on the claim to generate queries. The lack of background knowledge could cause hallucinatory or redundant queries that only address explicit aspects of the claim[85]. The huge gap in performance between the ORACLE-QUERIES and other methods further verifies this observation. Additionally, a known disadvantage of COMBMAX is that it sometimes tends to increase noise compared to COMBSUM which has a self-correcting nature[70]. Hence, a more balanced approach may be needed.

5.2.3 Key takeaways

In summary, the final takeaways from the evaluation of aggregation techniques and finally of key query planning methods are:

1. For aggregation of retrieval results across decomposed queries, COMBMAX-NORM delivers the highest performance, while TOP-1 performs the worst.
2. Evaluation of the retrieval performance of key query planning methods shows that decomposition-based query planning methods improve Recall@100, while NDCG@10 and Recall@10 scores remain comparable.
3. Additionally, we see that adding external signals to the retriever to reduce noise improves the overall performance of all query planning methods, especially those based on decomposition.

5.3 Downstream Impact

While the above results indicate the performance of the query planning methods at retrieval, observing their corresponding downstream impact is essential as it represents the final outcome of the pipeline. Hence, to answer **RQ3** we evaluate the final performance of the veracity prediction component for each of the query planning methods, the results of which are shown in Table.5.6.

Query mode	**NLI performance**					
	Accuracy	**F1-T**	**F1-F**	**F1-C**	**W-F1**	**M-F1**
ROBERTA-MNLI-CLAIM-NO-EV	63.2	24.71	79.93	47.37	61.53	50.67
ORACLE	65.25	47.76	79.92	49.67	66.56	59.11
CLAIM-ONLY	66.53	53.38	81.46	31.02	64.03	**55.28**
Decomposition						
ORACLE-QUERIES	66.45	45.73	80.78	50.25	66.8	**58.92**†
CLAIMDECOMP	64.57	53.78	79.97	35.81	64.41	56.51
PGMFC	65.67	51.38	81.19	37.15	65.73	56.57
QGEN	66.25	53.62	81.39	36.95	65.41	**57.28**†

Table 5.6: NLI Performance Metrics for Various Query Modes on QUANTEMP++.† indicates statistical significant with respect to CLAIM-ONLY at 0.05 level. Here W-F1 and M-F1 are weighted and macro F1 respectively. F1-T, F1-F, and F1-C represent the per-class F1 scores for the True, False, and Conflicting classes, respectively.

Superiority of decomposition-based methods. Firstly, we observe that the performance of ROBERTA-MNLI-CLAIM-NO-EV performs the worst, corroborating that solely relying on the surface pattern of the claim to predict a veracity label provides a performance that is nearly as poor as random guessing. Hence, augmenting the NLI with evidence serves as context or

clarification, thereby improving performance. Secondly, we observe that the downstream performance provided by decomposition-based query planning methods, specifically that of QGEN is higher than that of CLAIM-ONLY. While PGMFC and CLAIMDECOMP are minimally better than CLAIM-ONLY, QGEN provides a significant performance gain of 3.6% in the Macro-F1 score. A comparison of the per-class F1 scores shows that decomposition-based query planning methods excel at claims that are of a conflicting nature. We observe relative gains of up to 20% by PGMFC compared to CLAIM-ONLY. Claims whose veracity is of conflicting nature specifically require diverse perspectives to be retrieved as evidence since some parts of them may be true and other parts false. In Table.5.8 (b), we see an example of this where CLAIM-ONLY fails to provide the correct downstream result as the evidence it retrieves is too homogeneous.

Superior performance of QGEN. Another interesting observation we see is that, despite PGMFC and CLAIMDECOMP using a larger LLM GPT-3.5-TURBO for decomposition, the downstream impact of QGEN's query decomposition method, which was developed by training a smaller model of Flan-T5 with our oracle queries from QUANTEMP++, is superior. The training of QGEN can also be seen as a form of knowledge distillation, as our oracle queries were generated by prompting the GPT-3.5-TURBO model with the claim and justification document as input. Distillation of knowledge from larger LLMs to comparatively smaller LLMs has been an effective strategy to reduce inference costs while maintaining performance [62, 11, 45, 21]. Specifically, FLAN-T5-LARGE [38] has shown to outperform zero shot LLMs like GPT-3.5-TURBO at specific in-domain tasks [18]. Additionally, Wu et al. [79] has shown that problem decomposition tasks are easier to distill into small LLMs in QA tasks. We observe the same with decomposing claims into sub-queries, which aids in retrieving and verifying information.

Performance by Claim Taxonomy. To understand what type of numerical claims benefit most from different query planning methods, we show veracity prediction results per numerical taxonomy class in Table.5.7. We see that decomposition-based methods provide significantly superior performance for comparison and interval class of numerical claims. PGMFC provides relative gains of about 11.5% over CLAIM-ONLY for comparison-based numerical claims and QGEN provides gains as high as 19% for interval-based numerical claims. This is only natural because interval and comparison-based numerical claims require multiple independent aspects to be fetched, which are then reasoned about to get the veracity of the claim. In contrast to interval and comparison-based numerical claims, we find that in statistical and temporal classes, the performance across various query planning methods is comparable.

Taxonomy	Temporal		Statistical		Interval		Comparison	
Method	**Accuracy**	**M-F1**	**Accuracy**	**M-F1**	**Accuracy**	**M-F1**	**Accuracy**	**M-F1**
ORACLE	80.62	62.48	58.51	54.68	66.86	57.55	54.12	53.25
NAIVE-CLASSIFIER	74.09	28.35	50.83	0.22	63.4	25.86	32.94	16.51
ROBERTA-MNLI-CLAIM-NO-EV	77.68	51.7	57.77	46.62	61.96	46.36	51.76	45.64
CLAIM-ONLY	79.74	54.22	62.56	55.98	64.55	47.41	53.33	50.69
Decomposition								
ORACLE-QUERIES	78.27	54.7	61.16	56.29	67.15	54.9	59.22	58.23
CLAIMDECOMP	77.09	54.67	59.59	55.25	64.55	53.3	55.29	54.79
PGMFC	76.95	49.94	61.49	55.9	65.99	53.17	64.57	**56.51**
QGEN	78.41	51.65	61.49	56.53	67.15	**56.38**	55.29	54.79

Table 5.7: Performance metrics of different query planning methods across Different Taxonomies and Models on QUANTEMP++

(a) Error:	**Missed Explicit Aspect: Noisy Retrieval**
Claim	Jimmy Carter The Southern Baptist Convention voted 13 years ago "that women were inferior and had to be subservient to their husbands."
Oracle	**[1]** Former President Jimmy Carter Leaves the Southern Baptist ...**[2]** Southern Baptist Convention's declaration that wives should "submit graciously" to their husbands...**[3]** Southern Baptist Convention. In this study, a ... never tarnish the influence of the church.
Queries	**[Q1]** The Southern Baptist Convention voted 13 years ago. **[Q2]** The Southern Baptist Convention voted that women were inferior and had to be sub-servient to their husbands.
Retrieved	**[1]** Former President Jimmy Carter Leaves the Southern Baptist ...**[2]** SBC and Women Pastors - Seminary survey...churches where women are pastoring ... **[3]** An Aid To Understanding the SBC - Baptist Press This brief paper is offered to assist..
Comment	The claim talks about the SBC voting on women being inferior. The SBC only mentioned that wives should "submit graciously" to their husbands, but did not mention anything about equality. Due to under-specified queries in PGMFC, the retrieved results are not fully relevant
(b) Error:	**Missed Explicit Apect: Homogenous Results**
Claim	Rodrigo Duterte President Rodrigo Duterte says no country can diminish the importance of the July 2016 arbitral award"
Oracle	**[1]** Duterte stresses soft approach toward China in last policy speech Referring to the July 2016..**[2]**...Arbitral ruling can't be ignored by any country, Duterte to Asean MANILA...**[3]** ...Philippine President Duterte has downplayed the South China Sea Arbitration... Award..
Queries	Same as Claim.
Retrieved	**[1]**...Arbitral ruling can't be ignored by any country, Duterte to Asean MANILA...**[2]** ...The Philippines "vigorously pushed" for the inclusion of an arbitration ruling ... **[3]** Duterte stresses soft approach toward China in last policy speech
Comment	The claim mentions that Rodrigo Deuterte stressed the importance of the July 2016 arbitral award. While he did do so, it conflicts with his previous views where he has downplayed its importance. We see that only evidence that confirms the claim is fetched by CLAIM-ONLY.
(c) Error:	**Missed Implicit Aspect**
Claim	Jim Renacci "Since President Obama took office, our federal spending has increased by nearly 30 percent and our national debt has increased by almost 50 percent."
Oracle	**[1]** President Obama's Spending...Spending has gone up from $2.98 trillion in 2008 to...**[2]**...US debt: how big is it and who owns it? ...has gone up from $3tn, a rise of 74.1%...**[3]** ...In 2008...financial crisis had generated a decrease in government revenues and an increase in government expenditures... Award..
Queries	**[Q1]** Federal spending has increased by nearly 30 percent since President Obama took office.**[Q1]** The national debt has increased by almost 50 percent since President Obama took office
Retrieved	**[1]**...The amount of federal debt held by the public has skyrocketed in the past...**[2]** ...National Debt Increased Under Obama Faster Than Any Other... **[3]** Federal Government Will Pick Up Nearly All Costs of Health ...
Comment	While the numbers for spending and debt are correct, it was caused by factors out of Obama's control such as the financial crisis and recession. This implicit aspect was not fetched as the queries don't address it CLAIMDECOMP.

Table 5.8: Examples of Retrieval Error Codes recognized during analysis.

Discrepancy between retrieval and downstream performance. Lastly, an interesting observation that we see from comparing the results from retrieval in Table.5.5 and NLI from Table.5.6 is that the gains in retrieval performance do not proportionally translate to downstream NLI gains. The retrieval performance of oracle queries provides significant relative gains of 68.7% at Recall@10 and 74% at NDCG@10 over the query planning method of CLAIM-ONLY at retrieval. However, we see that this only translated to a relative gain of 6.6% downstream. According to Zamani et al. [82], relevance-based retrieval pipelines are designed with the assumption that retrieved information will be consumed by humans. However, this approach might not be suitable for retrieval-enhanced verification frameworks. To address this, Zhang et al. [83] optimizes the retriever by incorporating the utility of retrieval in downstream tasks as feedback during the retriever training process. Following the same philosophy to train the query decomposition model may reduce the observed gap in the performance of retrieval and downstream NLI.

5.3.1 Error Analysis

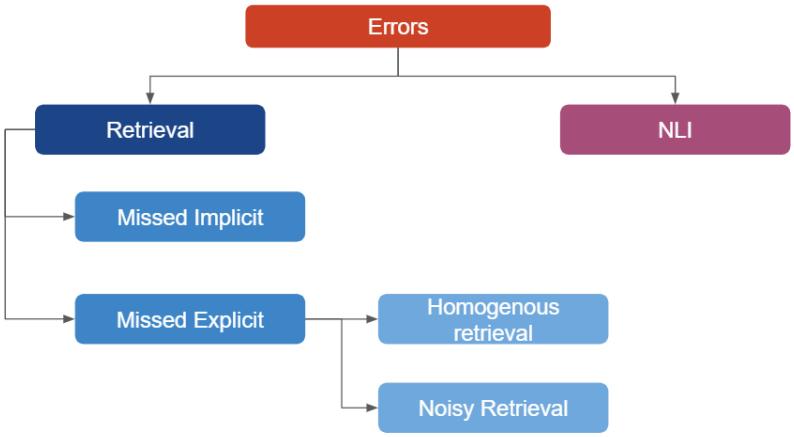


Figure 5.2: Taxonomy of errors identified during Error analysis

Error analysis of query planning methods helps us better understand their performances and identify their pitfalls. Our thematic analysis on 25 erroneous samples per query planning method helped us form a taxonomy of error codes provided in Fig.5.2. Table.5.10 and Table.5.2 show the distribution of errors per query planning method. Table.A.2 in the appendix additionally shows the distribution of errors per query panning method and claim veracity class. Table.A.1 in the appendix shows one example per error code observed.

Retrieval as the main source of error. We see that for all query planning methods, the primary source of errors are in the retrieval stage. This shows that it is essential to build a strong retrieval pipeline to improve the overall performance of the claim verification pipeline. In Table.5.10 we observed that within retrieval errors, two types of errors are caused. The retrieved evidence either fails to address explicit aspects of the claim or implicit aspects of the claim. An

Error Code	CLAIM-ONLY	CLAIMDECOMP	PGMFC	QGEN
Retrieval errors				
Missed Implicit Aspects	6	8	4	7
Missed Explicit Aspects	12	9	11	5
Unverifiable				
Evidence unavailable	3	4	3	0
Label error	0	1	1	3
NLI Errors	5	5	6	9

Table 5.9: Distribution of error Codes Across different query planning methods.

Error Codes	Claim Only	Claim Decomp	PgmFC	QGen
Homogenous retrieval	11	9	6	4
Noisy retrieval	1	0	5	1

Table 5.10: Distribution of Missed Explicit Errors across different query planning methods.

implicit aspect refers to an element of the claim that isn't explicitly stated but is inferred. Understanding these aspects often relies on background knowledge or on numerical common sense[9]. The majority of retrieval errors, however, are due to retrieved evidence failing to address all the multiple aspects explicitly mentioned in the claim. Table.5.8 shows examples for each of these cases.

Query Diversity and Granularity. On further analysis of errors caused by missed explicit aspects in retrieval (See Table.5.9), we see that they are caused by either homogeneous retrieval or noisy retrieval. Homogeneous retrieval occurs when all the retrieved evidence predominantly aligns closely or agrees entirely with the query. This phenomenon is expected in the CLAIM-ONLY setting where we provide only the claim to the retriever. Given there can be snippets with diverse perspectives on the claim from different sources, those with significant overlap with the claim are prioritized and pushed to the top. Decomposition-based query planning techniques such as CLAIMDECOMP and PGMFC, the presence of under-decomposed queries and redundant queries in the QGEN also contribute to this issue. Previous research such as Santos et al. [55] that enforce diversification of search results has shown to provide superior results. In the case of PGMFC, errors arise from noisy retrieval due to under-specification, and in case of QGEN, errors arise from hallucinated queries.

5.3.2 Key takeaways

1. Decomposition-based query planning methods provide superior downstream performance, especially for numerical claims of conflicting nature.
2. The ability to decompose numerical claims can be effectively distilled into smaller models, enabling them to outperform larger LLMs that rely on few-shot techniques
3. During query planning, over-specified queries lead to homogeneous retrieval results whereas under-specified queries lead to noisy results. Therefore, a balanced approach is essential.
4. Finally, the retrieval performance may not translate proportionally to downstream NLI performance as the relevance of retrieved evidence may not corroborate with their utility to verify

5.4 Ablations

5.4.1 Performance of Zero-Shot Retrievers

Table.5.11 shows the zero-shot retrieval performances of different retrievers. We see that BM25, CONTRIEVER, and ANCE perform comparably at the metric of NDCG@10 and Recall@10, but CONTRIEVER provides the best performance for Recall@100. This superior performance could be due to contrastive learning that helps retrieve relevant evidence in the presence of distractors. Similarly, ANCE which is also trained using contrastive learning performs comparably to CONTRIEVER while giving slightly lower recall values.

Zero Shot Retriever	NDCG@10	Recall@10	Recall@100
BM25	0.59	0.55	0.69
CONTRIEVER	**0.57**	**0.55**	**0.82**
DPR	0.41	0.37	0.67
TAS-B	0.53	0.49	0.79
ANCE	0.58	0.54	0.80

Table 5.11: Retrieval performance of ORACLE-QUERIES on the validation split of QUANTEMP++

5.4.2 NLI Input Length

In the fact verification pipeline, post retrieval, only a fixed number of top evidence snippets can be provided to the downstream MNLI model as most MNLI models have input length restrictions. The number of snippets we provide can have varying impacts on the downstream results. In Table.5.12, we see that all the metrics first increase from Top-1 to Top-3, then decrease and stagnate after. Providing too few snippets can impair the coverage of the aspects of the numerical

claim and including too many snippets can introduce noise in the set and again reduce performance. From Figure.5.3 we see that within the retrieval performance of ORACLE-QUERIES, recall increases with k (number of retrieved snippets), and precision increases from k=1 to k=3 then reduces.

Query mode	Accuracy	W-F1	M-F1
Top 1	64.57	62.74	54.28
Top 3	**66.45**	**66.80**	**58.92**
Top 5	66.25	65.20	56.90
Top 7	65.67	65.65	57.30
Top 10	65.65	65.30	57.30

Table 5.12: NLI performance of ORACLE-QUERIES with different input evidence lengths.

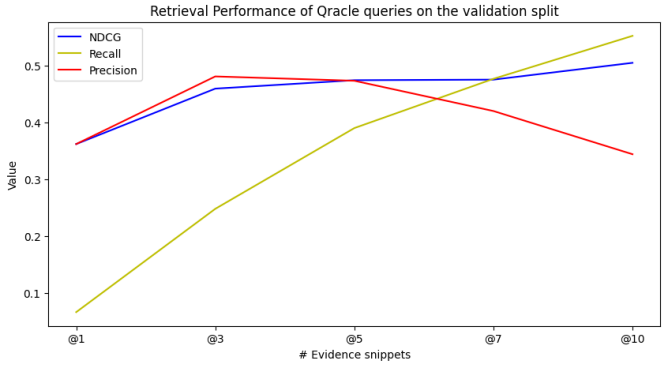


Figure 5.3: Retrieval performance of ORACLE-QUERIES on QUANTEMP++ on the validation set.

Chapter 6

Conclusion and Future Work

In our work, we aim to create a comprehensive dataset of natural numerical claims that provide a realistic environment to develop effective automatic fact verification pipelines that could possess the skills of a human fact checker. Given this realistic dataset, we also aim to evaluate the performance of existing query planning techniques, as they form the foundation of the fact verification pipelines and hence can significantly impact downstream performance. We assess the pros and cons of key existing query planning paradigms, particularly those of decomposition with respect to retrieval performance and then finally downstream veracity prediction performance. Through experiments, we addressed the following research questions:

1. Does query decomposition help retrieve quality evidence from the web for the verification of natural numerical claims?
2. How do existing query planning methods perform in terms of retrieval of relevant evidence snippets to verify numerical claims?
3. What is the downstream impact of these query planning methods on the task of verification of numerical claims?

On analysis, we find that our dataset, QUANTEMP++, generated using weak supervision effectively captures the implicit and explicit numerical aspects of the claim through decomposed queries. These queries then help retrieve evidence from the open web. Fine-tuning a smaller NLI model with our dataset provides superior performance over the baselines included. Surprisingly, we observe that a smaller NLI model fine-tuned with the claim and oracle evidence as input outperforms the larger GPT-3.5-TURBO model that utilizes a few shot settings, achieving relative gains of about 36%.

To evaluate the retrieval performance of key query planning paradigms, specifically those based on decomposition, we require a principled method to combine the results across the decomposed queries. We see that COMBMAX-NORM provides superior performance compared to other methods. Using COMBMAX-NORM, we then see decomposition-based query planning methods provide superior recall compared to solely using the claim for retrieval. Additionally, we see that adding external signals to the retriever to reduce noise improves the overall performance of all query planning methods, especially those based on decomposition. When evaluating the downstream impact of these methods, we consistently find that decomposition-based query planning methods outperform others. Notably, QGEN, a query generation model

trained with our oracle queries, achieves up to 3.6% relative gains over just using the claim as a query. This demonstrates that the ability to decompose numerical claims can be effectively distilled into smaller models. Specifically for interval and comparison-based claims, decomposition methods prove superior performance because of their ability to capture diverse perspectives. Finally, on comparing the retrieval performance of the oracle queries from our dataset and their downstream impact, we see that the retrieval performance may not translate proportionally to downstream NLI performance as the relevance of retrieved evidence may not corroborate with their utility to verify.

Overall, the choice of query planning method can largely impact retrieval and hence the final performance of the verification pipeline. Under-specified queries can introduce noise into the system, whereas over-specified queries can limit diversity, thereby hindering the pipeline's performance for natural numerical claims. Therefore, while claim decomposition shows promising results, it is essential to use a balanced approach to produce these decompositions. In essence, creating a realistic environment and focusing on refining the initial component of the pipeline can help address problems at the root, preventing error propagation and enabling the proper development of subsequent components. With our research, we aim to provide this realistic environment, allowing future researchers to understand the challenges of developing a fact verification pipeline suitable for real-time deployment for natural numerical claims. The development of such a pipeline can greatly assist journalism by managing the overwhelming number of numerical claims, helping to protect the public from deception, and potentially even saving lives.

6.1 Limitations and Future Work

While our research provides a realistic dataset for verifying natural numerical claims and provides meaningful insights into the query planning needed to address their information needs, we acknowledge the following drawbacks.

Data Creation Pipeline

With our data creation pipeline we assume that for each claim, the supporting evidence is publicly available on the web. However, in reality human fact-checkers also get their evidence from private repositories, websites behind paywalls, or by making calls to different institutions requesting information. Additionally, in our data pipeline, we consider a single modality-text, however, in reality, evidence can be in the form of videos, audio, images, and PDF documents[2]. Therefore, a more comprehensive approach is needed to gather evidence from multiple modalities and archived pages which would improve evidence coverage of the given claims. We additionally see that the qualitative analysis of the set of queries and evidence snippets generated during our data creation pipeline indicates that despite applying our filters, there is still a considerable amount of noise and redundancy. Hence a stronger filter using an appropriate unsupervised clustering mechanism[24] could be employed to only retain representative queries and evidence snippets.

Instillation of numerical sense

During the evaluation of downstream veracity classifiers, we saw that the performance of the models trained using justification documents produced a maximum Macro-F1 of 61% leaving a large gap for improvement. Similarly, in our evaluation of retrieval, we observe a large gap in performance between oracle queries and other paradigms indicating the lack of numerical understanding and reasoning skills in these components. Several works like GENBERT[20], NumBERT[73] and method proposed by Petrak et al. [46] have tried to inject numerical reasoning ability into language models and shown improvements in downstream tasks like discrete reasoning. Future works could assess their effectiveness in retrieval tasks and their superiority in verifying factual numerical claims.

Improving query generation

Comparing the retrieval and downstream performances of oracle queries in QUANTEMP++ showed us that retrieval performance did not translate proportionally to downstream veracity prediction performance. In a closely related study aimed at addressing this issue, Zhang et al. [83] enhances the retriever by integrating the utility of retrieval in downstream tasks as feedback during the retriever training process. Following the same philosophy to train the query decomposition model may reduce the observed gap in the performance of retrieval and downstream NLI. Furthermore, using reinforcement learning to explicitly reward the diversity and uniqueness of queries during the training of the query generator could also help mitigate other disadvantages of existing methods.

Utilization of full web pages

Finally, in our work, we utilize search snippets as evidence despite crawling the full-length web pages corresponding to these snippets. This restricts our work as full-length web pages provide full context and contain important information such as disclaimers and credibility characteristics that play an important role in the verification of natural claims. However, since most LLMs have limited input length, utilizing these full-length web pages is not as straightforward. A suite of multilevel retrieval, chunking, and aggregation techniques is required to process such information, which we leave for future researchers to explore.

Bibliography

[1] SerpApi: Google Search API — serpapi.com. https://serpapi.com/. [Accessed 08-06-2024].

[2] Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.361. URL https://aclanthology.org/2023.findings-emnlp.361.

[3] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5513. URL https://aclanthology.org/W18-5513.

[4] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*, 2021.

[5] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1475. URL https://aclanthology.org/D19-1475.

[6] Aditya Kiran Brahma, Srinivas Nagamalla, Jose Mathew, and Jairaj Sathyanarayana. Improving search relevance in a hyperlocal food delivery using language models. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data*

(11th ACM IKDD CODS and 29th COMAD), CODS-COMAD '24, page 479–483, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716348. doi: 10.1145/3632410.3632428. URL https://doi.org/10.1145/3632410.3632428.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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

[10] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. *ArXiv*, abs/2305.11859, 2023. URL https://api.semanticscholar.org/CorpusID:258822852.

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[12] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. Computation-al journalism: A call to arms to database research-ers. In *5th Biennial Conference on Innovative Data Systems Research, ACM*, 2011.

[13] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings*

of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 69–76, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2006. URL https://aclanthology.org/S17-2006.

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

[15] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. The effects of crowd worker biases in fact-checking tasks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2114–2124, 2022.

[16] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. Generating fact checking briefs. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.580. URL https://aclanthology.org/2020.emnlp-main.580.

[17] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23, 1981.

[18] Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization?, 2024.

[19] Carlo Galli, Nikolaos Donos, and Elena Calciolari. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. *Information*, 15(2):68, 2024.

[20] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.89. URL https://aclanthology.org/2020.acl-main.89.

[21] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models, 2024.

[22] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL https://aclanthology.org/2022.tacl-1.11.

[23] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, page 113–122, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462891. URL https://doi.org/10.1145/3404835.3462891.

[24] Yuan Hong, Jaideep Vaidya, Haibing Lu, and Wen Ming Liu. Accurate and efficient query clustering via top ranked search results. In *Web Intelligence*, volume 14, pages 119–138. IOS Press, 2016.

[25] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021. URL https://arxiv.org/abs/2112.09118.

[26] Pegah Jandaghi and Jay Pujara. Identifying quantifiably verifiable statements from text. In Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik, editors, *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pages 14–22, Toronto, ON, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.matching-1.2. URL https://aclanthology.org/2023.matching-1.2.

[27] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. Hover: A dataset for many-hop fact extraction and claim verification, 2020.

[28] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

[29] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

[30] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*, 2023.

[31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

[32] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

[33] Kashif Khan, Ruizhe Wang, and Pascal Poupart. WatClaimCheck: A new dataset for claim entailment and inference. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.92. URL https://aclanthology.org/2022.acl-long.92.

[34] Ryosuke Kinoshita and Shun Shiramatsu. Agent for recommending information relevant to web-based discussion by generating query terms using gpt-3. In *2022 IEEE International Conference on Agents (ICA)*, pages 24–29, 2022. doi: 10.1109/ICA55837.2022.00011.

[35] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99: 101861, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101861. URL https://www.sciencedirect.com/science/article/pii/S156625352300177X.

[36] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623.

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

[38] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[39] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466.

[40] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL https://aclanthology.org/2022.acl-long.244.

[41] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):258–267, Aug. 2021. doi: 10.1609/icwsm.v9i1.14625. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14625.

[42] Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. Multi-hop fact checking of political claims. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/536. URL https://doi.org/10.24963/ijcai.2021/536. Main Track.

[43] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*, 2023.

[44] Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. Faviq: Fact verification from information-seeking questions. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:235731930.

[45] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[46] Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. Arithmetic-based pretraining improving numeracy of pretrained language models. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 477–493, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.42. URL https://aclanthology.org/2023.starsem-1.42.

[47] Aman Rangapur, Haoran Wang, and Kai Shu. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation, 2023.

[48] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

[49] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google, 2019.

[50] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

[51] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, WWW '20, page 1160–1170, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380193. URL https://doi.org/10.1145/3366423.3380193.

[52] Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264, 2023.

[53] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.165. URL https://aclanthology.org/2021.acl-long.165.

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

[55] Rodrygo LT Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. Explicit search result diversification through sub-queries. In *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32*, pages 87–99. Springer, 2010.

[56] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web, 2023.

[57] Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*, 2021.

[58] Dhwanil Shah, Krish Shah, Manan Jagani, Agam Shah, and Bhaskar Chaudhury. Concord: Numerical claims extracted from the covid-19 literature using a weak supervision approach. *Available at SSRN 4222185*, 2023.

[59] Pratvi Shah, Arkaprabha Banerjee, Agam Shah, Bhaskar Chaudhury, and Sudheer Chava. Numerical claim detection in finance: A weak-supervision approach. *TechRxiv preprint*, 21288087, 2022.

[60] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.230. URL https://aclanthology.org/2021.emnlp-main.230.

[61] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[62] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[63] James Thorne and Andreas Vlachos. An extensible framework for verification of numerical claims. In André Martins and Anselmo Peñas, editors, *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-3010.

[64] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[65] Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims, 2024.

[66] Art Van Zee. The promotion and marketing of oxycontin: commercial triumph, public health tragedy. *American journal of public health*, 99(2):221–227, 2009.

[67] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22, 2014.

[68] Andreas Vlachos and Sebastian Riedel. Identification and verification of simple claims about statistical properties. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors,

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2596–2601, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1312. URL https://aclanthology.org/D15-1312.

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

[70] Christopher C Vogt and Garrison W Cottrell. Fusion via a linear combination of scores. *Information retrieval*, 1(3):151–173, 1999.

[71] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

[72] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347.

[73] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534.

[74] Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.416. URL https://aclanthology.org/2023.findings-emnlp.416.

[75] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067.

[76] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

[77] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating scientific claims for zero-shot scientific fact checking, 2022.

[78] Shengli Wu, Fabio Crestani, and Yaxin Bi. Evaluating score normalization methods in data fusion. In *Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006, Singapore, October 16-18, 2006. Proceedings 3*, pages 642–648. Springer, 2006.

[79] Zhuofeng Wu, He Bai, Aonan Zhang, Jiatao Gu, VG Vydiswaran, Navdeep Jaitly, and Yizhe Zhang. Divide-or-conquer? which part should you distill your llm? *arXiv preprint arXiv:2402.15000*, 2024.

[80] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval, 2020.

[81] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[82] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. Retrieval-enhanced machine learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2875–2886, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531722. URL https://doi.org/10.1145/3477495.3531722.

[83] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. From relevance to utility: Evidence retrieval with feedback for fact verification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.422. URL https://aclanthology.org/2023.findings-emnlp.422.

[84] Jian Zhang, Jianfeng Gao, Ming Zhou, and Jiaxing Wang. Improving the effectiveness of information retrieval with clustering and fusion. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 6, Number 1, February 2001: Special*

Issue on Natural Language Processing Researches in MSRA, pages 109–125, February 2001. URL https://aclanthology.org/O01-2005.

[85] Xuan Zhang and Wei Gao. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.64. URL https://aclanthology.org/2023.ijcnlp-main.64.

[86] Xin Zhou, Adrien Depeursinge, and Henning Müller. Information fusion for combining visual and textual image retrieval in imageclef@icpr. In Devrim Ünay, Zehra Çataltepe, and Selim Aksoy, editors, *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 129–137, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-17711-8.

Appendix A

Appendix

A.1 Examples

This section presents detailed examples of phenomena observed in the fact verification pipeline.

Error Code: Missed Explicit Apect: Homogenous Results
Example of method: CLAIM-ONLY
Claim: Rodrigo Duterte President Rodrigo Duterte says no country can diminish the importance of the July 2016 arbitral award by the Permanent Court of Arbitration on the South China Sea.
Oracle Evidence:

1. Duterte stresses soft approach toward China in last policy speech Referring to the July 2016..
2. ...Arbitral ruling can't be ignored by any country, Duterte to Asean MANILA..
3. 021/126 "...Philippine President Duterte has downplayed the South China Sea Arbitration Award in the hope of gaining China's infrastructure and ...

Queries:

1. Same as Claim

Fetched:

1. ...Arbitral ruling can't be ignored by any country, Duterte to Asean MANILA..
2. ... Philippines says dropping South China Sea ruling from ASEAN...The Philippines "vigorously pushed" for the inclusion of an arbitration ruling ...
3. Duterte stresses soft approach toward China in last policy speech...Referring to the July 2016 ruling by an international tribunal ...

Comment: The claim mentions that Rodrigo Deuterte stressed the importance of the July 2016 arbitral award. While he did do so, it conflicts with his previous views where he has downplayed its importance. We see that only evidence that confirms the claim is fetched.

Error Code: Missed Explicit Aspect: Noisy Retrieval
Example of method: PGMFC
Claim: Jimmy Carter The Southern Baptist Convention voted 13 years ago "that women were inferior and had to be subservient to their husbands.".

Oracle Evidence:

1. Former President Jimmy Carter Leaves the Southern Baptist ... The sad reality is that Jimmy Carter has been estranged from the Southern Baptist Convention for decades. He shifted his identification to another ...
2. Wifely submission: The SBC Resolution — The Christian Century Media coverage of the Southern Baptist Convention's declaration that wives should "submit graciously" to their husbands has been surprisingly extensive.

PDF the political and social attitudes of political and social attitudes of Southern Baptist clergy, a ... Southern Baptist Convention. In this study, a ... never tarnish the influence of the church.

Queries:

1. The Southern Baptist Convention voted 13 years ago.
2. The Southern Baptist Convention voted that women were inferior and had to be subservient to their husbands.

Fetched:

1. Former President Jimmy Carter Leaves the Southern Baptist ... The sad reality is that Jimmy Carter has been estranged from the Southern Baptist Convention for decades. He shifted his identification to another ...
2. SBC and Women Pastors - Baptist Press According to the Midwestern Seminary survey ten state conventions have churches where women are pastoring Southern Baptist churches – half of them in Virginia.
3. An Aid To Understanding the SBC - Baptist Press This brief paper is offered to assist those who wish to understand the Southern Baptist Convention, its work, perspectives, and governance, better. BAPTISTS IN..

Comment: The claim talks about the SBC voting on women being inferior. The SBC only mentioned that wives should "submit graciously" to their husbands, but did not mention anything about equality. Due to under-specified queries, the retrieved results are not fully relevant.

Error Code: Retrieval Error:Missed Implicit Aspect

Example of method: PGMFC, CLAIMDECOMP

Claim: Jim Renacci "Since President Obama took office, our federal spending has increased by nearly 30 percent and our national debt has increased by almost 50 percent."

Oracle Evidence:

1. President Obama's Spending — Cato at Liberty Blog Spending has gone up from $2.98 trillion in 2008—the year before Obama came into office—to a proposed $3.80 trillion in 2013. That is a 28%percent increase in ...
2. US debt: how big is it and who owns it? — News — theguardian.com Under President Obama's first term, that figure has gone up from $3tn, a rise of 74.1%. Under George W Bush, it went up too - by 85% over the whole two terms - ....

[PDF The effect of the economic and financial crisis on government ... In 2008 and 2009, the economic and financial crisis had generated a decrease in government revenues and an increase in government expenditures in terms of GDP.

Queries:

1. Federal spending has increased by nearly 30 percent since President Obama took office.
2. The national debt has increased by almost 50 percent since President Obama took office.

Fetched:

1. The Budget and Economic Outlook: An Update The amount of federal debt held by the public has skyrocketed in the past two years: from 40 percent of GDP at the end of 2008 to nearly 62 percent at the end ...
2. National Debt Increased Under Obama Faster Than Any Other ... An analysis by CBS News shows that the national debt has skyrocketed under President Barack Obama — increasing more than $4 trillion during his presidency.
3. Federal Government Will Pick Up Nearly All Costs of Health ... Congressional Budget Office (CBO) analysis indicates that between 2014 and 2022, the ACA's Medicaid expansion will add just 2.8 percent to what states spend on ...

Comment: The claim mentions that federal spending and debt have increased since Obama took office. While the numbers are correct, it was caused by factors out of Obama's control such as the financial crisis and recession. This implicit aspect was not fetched as the queries don't address it.

Error Code: Unverifiable: Evidence Unavailable

Example of method: PGMFC

Claim: Bill Pascrell "As many as 22,000 Americans die each year because they don't have health insurance." a speech

Oracle Evidence:

1. Health Insurance and Access to Health Care in the United States - Being uninsured is associated not only with inadequate access to care and poorer health but also with the most serious health consequence, premature death.
2. Why We Must Ration Health Care - The New York Times ... Research Institute, described ... care and had a death rate 37 percent higher than those with health insurance. ... Richard Kronick, a professor at the School of ...

PDF Uninsured and Dying Because of It: — Urban Institute insurance status and death rates. One used 1971 ... Americans died in 2000 because they were uninsured. ... "Mortality in the Uninsured Compared with that in.

3. More than 26,000 Americans die each year because of lack ... - NCBI In the seven years from 2000 to 2006 an estimated 162,700 Americans died because of lack of health insurance. Families USA said, "The number of uninsured ...
4. Health Insurance Coverage and Mortality Revisited - PMC - NCBI 1994;), and the IoM relied heavily on these two studies to estimate that lack of insurance increased the mortality rate by 25 percent. The studies were similar ...

Queries:

1. As many as 22,000 Americans die each year.
2. The cause of death for these Americans is the lack of health insurance.

Fetched:

1. Why We Must Ration Health Care - The New York Times Estimates of the number of U.S. deaths caused annually by the absence of universal health insurance go as high as 20,000. One study concluded that in the age ...
2. Health Insurance and Access to Health Care in the United States Being uninsured is associated not only with inadequate access to care and poorer health but also with the most serious health consequence, premature death.
3. More than 26,000 Americans die each year because of lack ... - NCBI Many more Americans die because of a lack of health insurance than previously thought, concludes a new state-by-state study by Families USA, a non-profit ...

Comment: The claim mentions that there are 22,000 Americans dying each year due to lack of insurance. While there is plenty of evidence confirming the same, a more recent paper by Richard Kronick of the University of California, published in HSR: Health Services Research, challenges the IOM's conclusions, making the final verdict conflicting. However, this new evidence is not present in the whole corpus itself, making it impossible for any retriever to fetch it.

Error Code: NLI Error

Example of method: QGEN

Claim: An aerial picture of burnt train cars in wet mud shows the aftermath of the February 2023 train derailment in East Palestine, Ohio.

Oracle Evidence:

1. Release of toxic chemicals from train derailment in Ohio prompts ... A drone footage shows the freight train derailment in East Palestine, Ohio. Ohio residents can return after air deemed safe after train derailment. By ... Railroad Lobbied Against Safety Changes Before Ohio Train.
2. Crash ... Ohio Train Derailment. Photo: mpi34/MediaPunch /IPX (AP). Last weekend, a cargo train carrying toxic chemicals derailed and caught fire in Ohio. Earlier this ... Ohio catastrophe is 'wake-up call' to dangers of deadly train ... Five days after a train carrying vinyl chloride derailed and exploded near the Ohio-Pennsylvania border, crews ignited a controlled burn of toxic chemicals to ...

Queries:

1. Aerial photograph train cars burning in wet mud East Palestine Ohio 2023
2. Atlanta International Airport train derailment February 2023
3. Train derailment aftermath East Palestine Ohio February 2023
4. Impact of train derailment on East Palestine Ohio community

Fetched:

1. File:2023 Ohio train derailment.jpg - Wikipedia Drone footage shows the freight train derailment in East Palestine, Ohio, U.S., February 6, 2023 in this screengrab obtained from a handout video released by ... East Palestine, OH Train Derailment Information. On the evening of February 3, 2023, a train derailed in East Palestine, Ohio. The tanker cars, carrying several types of chemicals, ruptured, and caught fire.
2. Site Profile - East Palestine Train Derailment - EPA OSC Response Please visit our East Palestine, Ohio Train Derailment Emergency Response webpage for the latest updates on EPA's emergency response to this incident.

Comment: Evidence clearly states that the incident mentioned by the claim did occur, yet the NLI model predicts Conflict as the label.

Table A.1: Examples of errors by code formed during error analysis.

Example: Claim from QUANTEMP++

Claim: `Donald Trump: Well, I got 52% (of women voters). And I will tell you | in my opinion I might have won women. ... Hey, Hillary Clinton focused on women, and I did phenomenally well..`

[Evidence Set 1]:

1. Donald Trump Claims He Won 52% Of White Women Vote In 2016 You know, I got 52% with women," President Donald Trump said at a press conference in late September. Subscribe to
2. Hillary Clinton 2016: How this presidential campaign will be different For starters, Hillary Rodham Clinton will emphasize her gender and women's issues more than she did in her 2008 presidential campaign.
3. Media Charged With Sexism in Clinton Coverage Angered by what they consider sexist news coverage of Senator Hillary Rodham Clinton's bid for the Democratic presidential nomination, many women and ...

[Verdict]: True

[Evidence Set 2]:

1. Donald Trump Claims He Won 52% Of White Women Vote In 2016 You know, I got 52% with women," President Donald Trump said at a press conference in late September...
2. The Gender Gap in Voting: Setting the Record Straight In the 2016 election, men were 11 percentage points more likely than women to vote for Donald Trump (52% of men vs. 41% of women), according to the exit ...
3. White Women Helped Elect Donald Trump - The New York Times While black and Hispanic or Latino women voted overwhelmingly for Hillary Clinton, 53 percent of white women who voted picked Mr. Trump, exit data show...

[Verdict]: False

Figure A.1: Example claim from QUANTEMP++ where different retrieved evidence can lead to different downstream veracity label

A.2 Prompts

AGQ Prompt

[CLAIM]
Prime Minister Narendra Modi breached the election protocol by addressing a rally in Howrah on April 6.
[END]

[PASSAGE]
Modi addressed a rally in Cooch Behar and Howrah's Dumurjola on April 6, where polls were held on April 10. The silence period was not breached. As the polling for the third phase of the West Bengal assembly election was underway for 31 seats on April 6, ...*skipped text*... In Dumurjola, the voting took place in the fourth phase on April 10. Therefore, Modi did not breach the 48 hours silence period. The voting for West Bengal assembly elections for 294 seats is taking place in eight phases from March 27 to April 29. The counting of votes would take place on May 2.
Published: 2021-04-07.
[END]

[QUERIES]
1. Modi rally in Howrah 2021
2. Prime Minister Narendra Modi rally on April 6
3. election in Howrah 2021
4. locations of Howrah voting
5. CM Mamata Banerjee Modi's rallies in 2021
6. silence period affects campaigning and media coverage of elections

[END]

Using the above as an example, generate at most 10 independent, short, and diverse Google search phrases required to verify or debunk the below claim labeled under [CLAIM]. Note: Generate diverse questions tending to the different numerical and temporal aspects both implicit and explicit to the claim.
Note: Use the passage under [PASSAGE] for reference to generate queries.
Do not generate queries for the passage.
Note: You must exclude any questions about fact-checking.

Figure A.2: Prompt used to generate sub-queries given claim and justification document in the data creation pipeline of QUANTEMP++

A.3 Instructions

Manual Annotation Instruction

In a fact verification process, a claim needs to be labeled as TRUE, FALSE or conflicting. To verify this claim a typical human verifier would perform a few web searches to collect evidence required to verify the claim. We aim to evaluate the quality of our dataset by evaluating the quality of the web search queries generated and the quality of the retrieved evidence snippets required to verify a given claim. This is carried out using two tasks.

Task 1
In sheet 1, you are given 50 claims and their corresponding web search queries. For a given claim, determine if each query is relevant or irrelevant. Tick the checkbox if the query is relevant. Additionally, for each claim rate the completeness and redundancy of the given web search queries on a scale of 1 to 5

The completeness rating could be interpreted in the following way:
5 - The set of queries cover all required aspects to verify the claim
4 - The set of queries cover most of the required aspects to verify the claim
3 - The set of queries cover some of the aspects required to verify the claim
2 - The set of queries cover miss most of the aspects required to verify the claim
1 - The set of queries cover no aspects required to verify the claim

The redundancy rating could be interpreted in the following way:
5 - The whole set of queries cover the same aspect needed to verify the claim
4 - Most of the queries cover the same aspect
3 - Some of the queries overlap with respect to covering an aspect
2 - Most of the queries adress a different aspect needed to verify a claim
1 - All queries cover indepent aspects required to verify a claim

Task 2
In sheet 2, you are given 50 claims and their corresponding evidence snippets. For a given claim, determine if each snippet is relevant or irrelevant. Tick the checkbox if the snippet is relevant. Additionally, for each claim rate the completeness and redundancy of the given snippets on a scale of 1 to 5. The meaning of the ratings for the evidence snippets is the same as it is for the queries related to the claim as mentioned above.

You would also be given the justification document used by a human fact verifier for reference. A summary of it generated by GPT-3.5 is also available

Two example annotations are shown below in the sheet ”Example”

Note that the claims across Task 1 and Task 2 are the same

Figure A.3: Instruction provided to annotators to assess the data quality of QUANTEMP++

A.4 Supplementary Tables

Errors by Class	Claim Only	Claim Decomp	PgmFC	QGen
FALSE	1	9	3	3
TRUE	4	2	1	3
CONFLICTING	20	14	21	19

Table A.2: Distribution of error codes by query planning method and veracity label identified during error analysis.

A.5 Model Summary

```
MultiClassClassifier(
  (roberta): RobertaModel(
    (embeddings): RobertaEmbeddings(
      (word_embeddings): Embedding(50265, 1024, padding_idx=1)
      (position_embeddings): Embedding(514, 1024, padding_idx=1)
      (token_type_embeddings): Embedding(1, 1024)
      (LayerNorm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (encoder): RobertaEncoder(
      (layer): ModuleList(
        (0-23): 24 x RobertaLayer(
          (attention): RobertaAttention(
            (self): RobertaSelfAttention(
              (query): Linear(in_features=1024, out_features=1024, bias=True)
              (key): Linear(in_features=1024, out_features=1024, bias=True)
              (value): Linear(in_features=1024, out_features=1024, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): RobertaSelfOutput(
              (dense): Linear(in_features=1024, out_features=1024, bias=True)
              (LayerNorm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
          (intermediate): RobertaIntermediate(
            (dense): Linear(in_features=1024, out_features=4096, bias=True)
            (intermediate_act_fn): GELUActivation()
          )
          (output): RobertaOutput(
            (dense): Linear(in_features=4096, out_features=1024, bias=True)
            (LayerNorm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
    (pooler): RobertaPooler(
      (dense): Linear(in_features=1024, out_features=1024, bias=True)
      (activation): Tanh()
    )
  )
  (dropout): Dropout(p=0.3, inplace=False)
  (mlp): Sequential(
    (0): Linear(in_features=1024, out_features=768, bias=True)
    (1): ReLU()
    (2): Linear(in_features=768, out_features=3, bias=True)
  )
)
```

Figure A.4: Summary of the MNLI model used to form our veracity classifiers.