# Contrastive Learning to Improve Retrieval for Real-world Fact Checking

**Aniruddh Sriram** **Fangyuan Xu** **Eunsol Choi** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
aniruddh.sriram@utexas.edu

## Abstract

Recent work on fact-checking addresses a realistic setting where models incorporate evidence retrieved from the web to decide the veracity of claims. A bottleneck in this pipeline is in retrieving relevant evidence: traditional methods may surface documents directly related to a claim, but fact-checking complex claims requires more inferences. For instance, a document about how a vaccine was developed is relevant to addressing claims about what it might contain, even if it does not address them directly. We present Contrastive Fact-Checking Reranker (CFR), an improved retriever for this setting. By leveraging the AVeriTeC dataset, which annotates subquestions for claims with human written answers from evidence documents, we fine-tune Contriever with a contrastive objective based on multiple training signals, including distillation from GPT-4, evaluating subquestion answers, and gold labels in the dataset. We evaluate our model on both retrieval and end-to-end veracity judgments about claims. On the AVeriTeC dataset, we find a 6% improvement in veracity classification accuracy. We also show our gains can be transferred to FEVER, ClaimDecomp, HotpotQA, and a synthetic dataset requiring retrievers to make inferences.

## 1 Introduction

Retrieval-augmented generation (RAG) systems are now widely used across NLP applications including question answering (Guu et al., 2020; Lewis et al., 2020; Karpukhin et al., 2020) and text generation (Komeili et al., 2022; Gao et al., 2023b), but one particular application of interest is fact-checking. While older fact-checking systems would often not consider evidence at all (Alhindi et al., 2018) or consider oracle evidence (Atanasova et al., 2020), the real fact-checking task involves finding evidence to support or refute complex claims in the wild (Chen et al., 2022; Schlichtkrull et al., 2023; Chen et al., 2024). As with many other RAG settings, retrieval is a bottleneck (Singh et al., 2022): it is impossible to provide the right judgment without retrieving the right evidence.

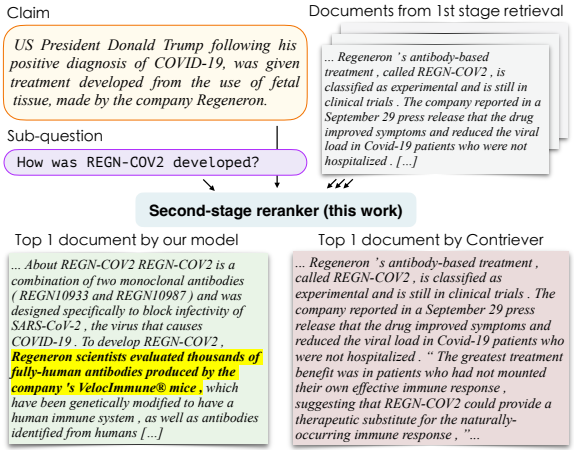


Figure 1: Top-1 retrieved document from base Contriever (red) and CFR (green). Our model is able to choose a better document despite both paragraphs being topical. Our model recognizes the question is asking about the chemical composition of REGN-COV2, while the unfinetuned model selects a relevant document that does not address "fetal tissue" or help with a final veracity judgment.

In this work, we investigate how to build an effective retriever for fact-checking. Figure 1 shows an example of why this is particularly challenging: unlike a factoid question with a definite answer spelled out in text, documents retrieved for fact-checking may only obliquely address a claim, or may present information in a different context (e.g., statistics that apply to a different country than the one where the claim was made). The unstructured nature of documents in the wild combined with claims that are only subtly true or false make retrieval a very difficult task.

We focus on two-step retrieval pipeline used in past work (Lazaridou et al., 2022; Chen et al., 2024). These use a first-stage web search (i.e., using Google or Bing) to build a set of approximately



*Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 264–279

relevant documents, followed by a second-stage fine-grained ranking to obtain a smaller set of documents to pass to a reader LM (Chen et al., 2024), which produces the final veracity judgment. This second stage shows consistent recall failures despite high-quality documents being present in the first stage, mainly due to the nuanced complexities with claims and subquestions in fact-checking.

Our approach, Contrastive Fact-Checking Reranker (CFR), leverages contrastive learning to fine-tune a dense retriever to prefer more relevant documents when there is a lack of information or ambiguity in the claim. To train our model, we experiment with two main supervision signals: distilling knowledge from GPT-4 and measuring answer equivalence with the gold answer using Learned Equivalence Metric for Reading Comprehension (LERC) (Chen et al., 2020). We generate training datasets of positive and negative evidence pairs based on these signals and fine-tune Contriever (Izacard et al., 2022).

Our evaluation shows that a combination of these supervision signals provides the best training data for the retriever, even better than fine-tuning on human annotated gold documents, as shown by gains in downstream performance across multiple datasets. Specifically, we see a 6% improvement in veracity classification accuracy and a 9% increase in the proportion of relevant top documents on AVeriTeC.

Our contributions are: (1) exploring new methods of supervision signals for contrastively training dense retrievers; (2) producing a strong dense retriever (CFR) which works well on AVeriTeC and a broader set of retrieval tasks regarding fact-checking complex claims.

## 2 Background and Related Work

### 2.1 Retrieval Augmented Generation Systems

Retrieval-augmented generation (RAG) relies on two key modules: a retriever and a reader/generation model. For many RAG systems, noisy retrieval hurts downstream performance by providing irrelevant or misleading documents (Yoran et al., 2024). Sauchuk et al. (2022) found that adding distractors can cause a 27% drop on veracity classification accuracy on FEVER. Therefore, it's important for retrievers to find relevant documents and simultaneously avoid damaging ones. Shi et al. (2023) attempts to solve this problem by finetuning the retrieval component while fixing the reader LM, similar to our work. Other approaches like Ke et al. (2024) create a more complex system with a "bridging" model between the retriever and reader. Nevertheless, noisy retrieval remains a failure point in RAG systems (Barnett et al., 2024), and tangible downstream gains can be realized by further finetuning.

### 2.2 Limitations of Existing Retrieval Systems

For NLP tasks like question answering, sparse retrieval techniques like BM25 have been supplanted by dense retrievers like DPR (Karpukhin et al., 2020) and Contriever (Izacard et al., 2022). These dual encoder approaches support efficient retrieval, and contrastive training is an effective way to learn embeddings for QA tasks. More recently, research has explored distilling knowledge from reader models to create smarter retrievers (Izacard and Grave, 2022). We draw from this work to build a retrieval system with better reasoning capabilities than baseline dense retrievers, which are usually pretrained on simpler (query, document) pairs (i.e. the MS-MARCO dataset). These retrieval systems have proven effective for fact-checking settings such as FEVER (Thorne et al., 2018) and MultiFC (Augenstein et al., 2019). However, the claims are largely short and factoid, and most of them contain no more than two entities. The realistic setting is embodied by approaches like QABriefs (Fan et al., 2020), ClaimDecomp (Chen et al., 2022, 2024), and AVeriTeC (Schlichtkrull et al., 2023), which are ultimately different from what dense retrievers were developed and optimized for.

### 2.3 Motivating Example: AVeriTeC

Figure 1 shows an example of fact-checking in the AVeriTeC dataset: *"how was REGN-COV2 developed?"*. This example differs in key ways from frequently-studied question answering settings such as such as Natural Questions (Kwiatkowski et al., 2019). First, it supports several different short answers but very likely has a best answer in the context of the claim: did the development involve human fetal tissue? In this case, the bolded paragraph indicates no: it used mice. The answer to this question should address the claim and provide background information: there is both a "short answer" as well as a "long answer" (Kwiatkowski et al., 2019; Gao et al., 2023a).

**Retrieval signals in fact-checking** Contrastive methods like Contriever require examples marked

as positive or negative for use in the contrastive objective. In settings like NQ, retrieval systems rely on evaluating whether a retrieved passage contains the answer by simple string matching or ROUGE overlap, which identifies "positives" for retrieval. However, in Section 5 we show it is not straightforward to apply this approach in fact-checking; i.e., we cannot simply say a passage is positive if it contains the ground truth answer.

Simultaneously, we must be cautious of assuming a low overlap with the answer indicates a "negative" document for retrieval. This is because multiple plausible answers can exist due to the open-ended nature of subquestions in AVeriTeC. Furthermore, using documents from the wild exacerbates this issue by introducing documents that might not directly support the gold answer but still contain valuable information about the claim. In Section 3, we outline some ways in which we tackle this problem to curate better finetuning data.

**Context in retrieval** Traditionally, retrievers are given standalone questions as queries. This is characteristic of datasets like NQ, where questions often contain one clear answer (e.g. "*Where is the bowling ball hall of fame located?*"). However, in fact-checking, the complexity of claims gives rise to subquestions that are not standalone or simple. Even if the questions themselves seem short (i.e., "*How was REGN-COV2 developed?*"), they must be interpreted in-context with the claim (i.e., "*Does REGN-COV2 contain fetal tissue?*"). Ideally, decomposing claims into a set of perfect standalone subquestions would reduce the load on the retriever. However, this itself is a hard and separate task. In this work, we attempt to build a retrieval system that can handle nuanced queries by considering each subquestion in the context of the overall claim.

## 3 Methodology

We consider a setting following work in AVeriTeC and ClaimDecomp (Chen et al., 2022). We assume we are given a collection of **claims** $(c_1, \ldots, c_N)$. For claim $c_i$, we define $q_{ij}$ as the $j$th **subquestion** for the $i$th claim in the dataset and $a_{ij}^g$ define its **answer**. We also assume access to a document set $D(c_i, q_{ij})$ for each subquestion, created by querying Bing with $c_i$ appended to $q_{ij}$ and scrape the top-k articles to form a document corpus. Each **document** $d$ is a 200 token span gathered from the scraped articles. The title of the document is prepended to the start of each document. The dataset also comes with a **gold article** which contains the gold answer. Like the Bing-retrieved documents, it is chunked into 200 token span documents $\{d^g\}$ and added to $D(c_i, q_{ij})$. We refer to documents belonging to these articles as *gold*.

Given a query $y = [c_i; q_{ij}]$ and a document $d_i \in D$, we want to generate embeddings in $\mathbb{R}^e$ using an encoder network (e.g. Contriever). Let $h_y, h_{d_i}$ denote the representations of $y$ and $d_i$. Then we define our scoring function $f : \mathbb{R}^e \times \mathbb{R}^e \to \mathbb{R}$ such that $f(h_y, h_{d_i}) > f(h_y, h_{d_j})$ if document $d_i$ contains more information helpful to answering the query than document $d_j$. Let $r(y) = \arg\max_{d \in D} f(h_y, h_d)$ which is a function that chooses the highest ranked document in our document set $D$. The goal is to optimize our encoder via $f$ to rank documents for answering questions in-context with the claim above topically relevant documents that do not ultimately contain information for an answer. We choose to optimize this for downstream veracity classification accuracy. We also track more upstream metrics such as using a relevance score for the top document or measuring how close its extracted answer matches the gold answer.

### 3.1 Components

**Dense retriever** $r$ We use Contriever as the base for our second stage dense retriever. Contriever uses the BERT base uncased architecture (Devlin et al., 2019). To fine-tune it with contrastive learning, we require document sets $T(c_i, q_{ij}, D) = \{D^+, D^-\}$ of positive and negative documents; during optimization, the positive documents will be embedded closer to the query vector than negative documents. Contrastive training relies critically on having hard negatives to serve as "distractors" (Robinson et al., 2021). These might be documents ranked high by baseline retrievers or having high token overlap with the query. Figure 2 shows our pipeline for constructing these document sets, which we expand on in the following sections.

We define $S_{\text{BM25}}(c_i, q_{ij}) = \{d_1, d_2, \ldots, d_k\}$ as the top $k$ documents surfaced by BM25 given $[c_i; q_{ij}]$ as the query. We also define $G_{\text{BM25}}(c_i, q_{ij}) = \{d_1^g, d_2^g, \ldots, d_l^g\}$ as the top $l$ gold annotated documents. In our models, we set $k = 10$ and $l = 5$.

**Reader model** We use GPT-4 as the reader model. The answers are derived by prompting GPT-4 with the claim $c_i$, question $q_{ij}$, and a document

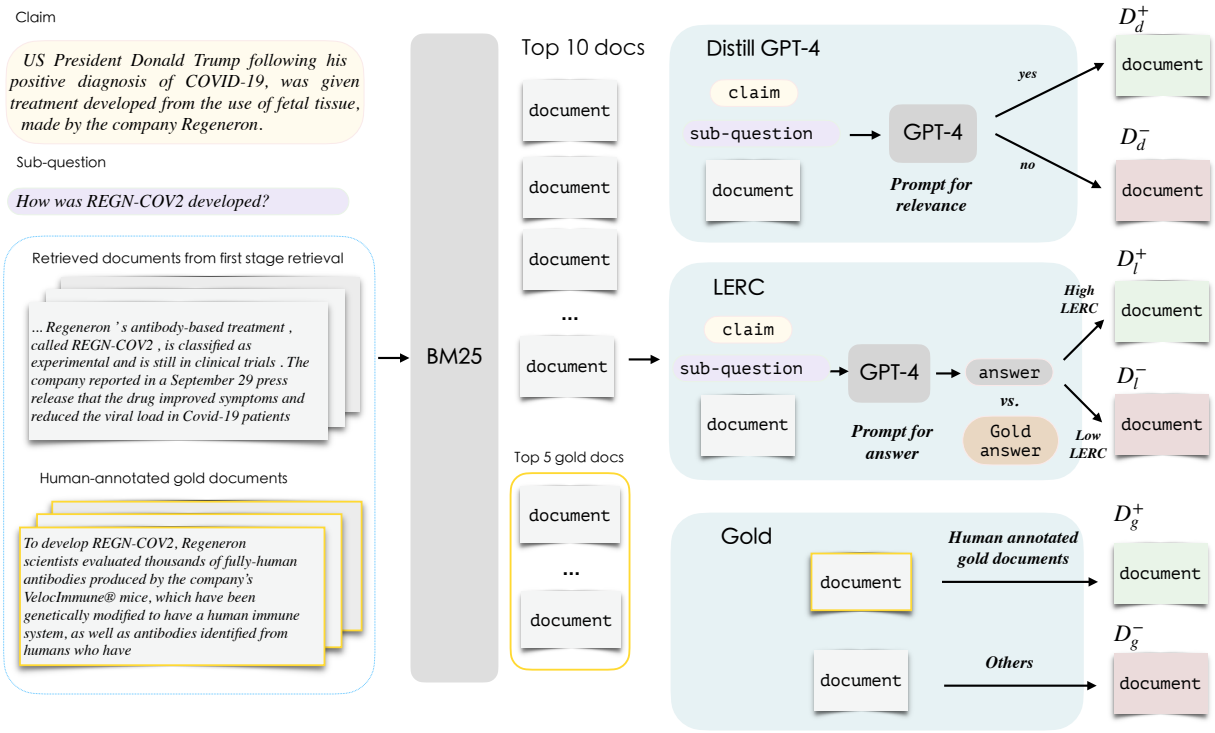


Figure 2: Overview of generating positive and negative examples for finetuning the retriever. We first select documents with high BM25 score with the (query, subquestion) from both the web documents and gold articles. We then experiment with different methods (described in Section 3.3) to derive positive and hard negative examples.

$d_{ij}$ from the corpus (see Appendix E.3). For a given $(c_i, q_{ij})$ pair, we refer to $a_{ij}$ as the candidate answer derived from the evidence document $d_{ij}$. During inference time, $d_{ij}$ is the top-1 document from our retrieval system.

### 3.2 Learning

We train $r$ on these $(c_i, q_{ij}) \times T$ pairs to produce a finetuned retriever $r^*$. Specifically, given a query $y = [c_i; q_{ij}]$ and positive document $d^+ \in D^+$,

$$L(y, d^+) = \frac{\exp\left(\frac{1}{\tau} f(h_y, h_{d^+})\right)}{\exp\left(\frac{1}{\tau} f(h_y, h_{d^+})\right) + \sum_{d^- \in D^-} \exp\left(\frac{1}{\tau} f(h_y, h_{d^-})\right)}$$

where $\tau$ is a temperature parameter. In our setting, we define $f$ as cosine similarity $\frac{h_y^T h_d}{\|h_y\| \cdot \|h_d\|}$ between the embeddings. This encourages positive documents to have high similarity with the query while penalizing high scores for negative documents. Fine-tuning yields $r^*$ such that $r^*(y)$ contains a better answer to $q_{ij}$ in context with $c_i$ than $r(y)$.

**Implementation Details** On average, each question $q_{ij}$ comes with about 500 documents to rank. Each document contains 200 token span, scraped from articles with a 100 token length stride. Details about training and model architecture can be found in Appendix A.1.

### 3.3 Generating Contrastive Training Data

We generate $\{D^+, D^-\}$ in three main ways: the annotated AVeriTeC gold evidence, distilled relevance judgements from a GPT-4 reader module, and evaluating equivalence of the document-predicted answer with a gold answer. Figure 2 shows the three approaches which we describe next.

**AVeriTeC Gold Evidence** The most straightforward approach to building positive examples is to use the human-annotated evidence paragraphs available in AVeriTeC. The gold articles (one per subquestion) were selected by human annotators in a two-stage annotation process, we refer the readers to their paper for details (Schlichtkrull et al., 2023). The annotators also provided answers for the subquestions, which consist of both extractive and abstractive answers. For each $q_{ij}$, this article is chunked into a set of documents $\{d_{ij}^g\}$ as described in Section 3. Negative examples are all $d \in S_{\text{BM25}}(c_i, q_{ij})$ such that $d$ is not from a gold-annotated document. We denote the fine-tuning data derived from this method as $\{D_g^+, D_g^-\}$.

**Distilling GPT-4** The AVeriTeC gold evidence may have recall errors: there may be relevant documents that are not marked by annotators. An al-

ternative is to use GPT-4 as a labeler, effectively distilling its knowledge (Figure 2, top right). In this setting, we take $S_{\text{BM25}}(c_i, q_{ij})$ and zero-shot prompt GPT-4 about whether each document is relevant to answering the subquestion or not. Note we do not provide the gold answer in the prompt, as we are simply interested in collecting documents with relevant information regardless of how well the underlying answer matches $a^g_{ij}$. Documents marked as relevant are added to $D^+$, and the rest are added to $D^-$. The exact prompt can be found in Appendix E.1. We define this set as $\{D^+_d, D^-_d\}$.

**Distilling GPT-4 (with gold)** In this setting, we inject the top-$l$ AVeriTeC gold documents $G_{\text{BM25}}(c_i, q_{ij})$ into the finetuning set. Like before, we zero-shot prompt GPT-4 about whether each document is relevant to answering the subquestion, but include $G_{\text{BM25}}(c_i, q_{ij})$ in addition to $S_{\text{BM25}}(c_i, q_{ij})$. We refer to $\{D^+_{dg}, D^-_{dg}\}$ as the finetuning data from this method.

**LERC-based signal** An additional approach to construct our pairs is to use the gold-annotated answers $a^g_{ij}$ (Figure 2, middle right). Ideally, a document we retrieve should help us discover these answers; however, because the subquestions are not factoid questions, it is not easy to assess whether a retrieved document contains the answer.

To do this, we filter the top documents using LERC (Learned Evaluation Metric for Reading Comprehension) (Chen et al., 2020), a metric for scoring answer equivalence. More formally, we take $S_{\text{BM25}}(c_i, q_{ij})$ with $G_{\text{BM25}}(c_i, q_{ij})$ to make a set of 15 documents. We then prompt GPT-4 to use each of the 15 evidence documents to produce an answer $a_{ij}$ for each document. We found that for complex long answers, using ROUGE overlap as an answer equivalence metric works poorly (Appendix B.1). On AVeriTeC, we also tried using ROUGE-F1 score instead of LERC (see Table 2) to see how this reflects in all our end-to-end evaluation metrics. To accommodate this, we introduce an "answer shortening" function $s$ which attempts to pull out the main point of the answer. We use LERC to compare $s(a_{ij})$ and $s(a^g_{ij})$, our shortened candidate and gold answer respectively. By identifying documents which give rise to answers with high LERC scores, we encourage our retriever to seek documents which address the question in the query. Documents with poor LERC scores ($< 0.3$) become negative contexts, and documents with high LERC ($> 0.7$) scores are positive contexts. We also evaluate how well human annotators agree with granular LERC scores and find an average Kendall's $\tau$ score of 0.53 (Appendix C.2). We denote $\{D^+_l, D^-_l\}$ as finetuning data derived from this method.

| Train Set | # subq | $\lvert D^+\rvert$ | $\lvert D^-\rvert$ | $D^+$ | $D^-$ |
|---|---|---|---|---|---|
| distill | 1228 | 4.8 | 8.4 | $D^+_d$ | $D^-_d$ |
| LERC | 692 | 1 | 4.2 | $D^+_l$ | $D^-_l$ |
| gold | 1229 | 1 | 9.1 | $D^+_g$ | $D^-_g$ |
| distill (gold) | 1229 | 5.2 | 8.4 | $D^+_{dg}$ | $D^-_{dg}$ |
| distill (gold) + LERC | 1229 | 5.6 | 8.4 | $D^+_{dg} \cup D^+_l$ | $D^-_{dg}$ |

Table 1: Dataset statistics for different finetuning sets from AVeriTeC. $\lvert D^+\rvert$ and $\lvert D^-\rvert$ represent the average number of positive and negative contexts per $(c_i, q_{ij})$ pair. Differences in number of subquestions come from filtering out examples for which $\lvert D^+\rvert = 0$ or $\lvert D^-\rvert = 0$.

**LERC-based quality check** We evaluated $\{D^+_l, D^-_l\}$ and found that many negative documents were actually relevant to the claim/question. More details on this experiment can be found in Appendix C.1. To reduce the false negative rate, we mix in relevant documents with the positive set from *distill* to create $\{D^+_{dg} \cup D^+_l, D^-_{dg}\}$. We refer to this as the *distill (gold) + LERC* setting. This is the final experimental setting we use for our **Contrastive Fact-Checking Reranker (CFR)** model.

## 4 Experimental Setup

We evaluate Contriever fine-tuned on the supervision signals outlined in Section 3. The datasets selected for evaluation, namely AVeriTeC (Schlichtkrull et al., 2023), ClaimDecomp (Chen et al., 2022), FEVER (Thorne et al., 2018), and HotpotQA (Yang et al., 2018), encompass a wide range of scenarios for document retrieval. For evaluation, a random subset of 200 answerable examples (subquestions contain an answer) were selected from each of these not overlapping with the training sets.

### 4.1 Metrics

We use metrics that evaluate both the retrieved documents and downstream products of these documents, such as the produced answer.

- **LERC** computes the average LERC score between the AVeriTeC (or ClaimDecomp) gold

answer and the GPT-4 generated answer from the top retrieved document as the candidate.

- **Top doc relevance** is the proportion of examples for which the top-1 document is classified as relevant to answering the question by GPT-4, using the same prompt for which we derive the distillation signal.
- **Gold@10** is the proportion of examples in which an AVeriTeC annotated gold document appeared in the top-10.
- **Veracity** represents the veracity classification accuracy. For ClaimDecomp, we use the RoBERTa based veracity classifier trained on ClaimDecomp.[1] For FEVER, we few-shot prompt GPT-4 for a veracity label; see Appendix E.4.

### 4.2 Datasets

**AVeriTeC** consists of real claims ($c_i$) from the web annotated with subquestions ($q_{ij}$), gold answers ($a^g_{ij}$) to the subquestions, and the gold evidence document for the answer. We query Bing in FSR with the claim and subquestion $[c_i; q_{ij}]$ to generate $D$. The generated answers ($a_{ij}$) are verified against the gold answers using LERC.

**ClaimDecomp** consists of complex political claims ($c_i$) with yes/no subquestion decompositions ($q_{ij}$) generated by trained annotators. We query Bing in FSR with the claim and subquestion $[c_i; q_{ij}]$ to generate $D$. The annotated subquestions tackle both explicit and implicit parts of the original claim. The implicit questions are much harder to answer without sufficient context, which makes this an interesting dataset for retrieval evaluation. The human labeled answers are yes/no, and we evaluate our generated answers ($a_{ij}$) against the gold answers using LERC. Because the questions themselves are yes/no in nature, this approach returns the same results as simple binary comparison.

**FEVER** consists of claims ($c_i$) manually verified against the introductory sections of Wikipedia pages and classified as SUPPORTED, REFUTED or NOTENOUGHINFO. We treat the claim itself as the question ($c_i = q_i$) here. Unlike past work, we query Bing with the claim to generate $D$; as a result, our data condition is different than past work evaluating on FEVER. For FEVER, we don't generate answers or subquestions and simply verify the claim against the evidence document.

[1] https://github.com/jifan-chen/Fact-checking-via-Raw-Evidence

| Model | LERC | Top Doc Relv. | Gold@10 | Veracity |
|---|---|---|---|---|
| BM25 | 0.45 | 0.47 | 0.42 | 0.48 |
| Contriever | 0.48 | 0.54 | 0.50 | 0.54 |
| Contriever MSM | 0.52 | 0.55 | 0.45 | 0.59 |
| ROUGE-F1* | 0.52 | 0.53 | 0.50 | 0.55 |
| gold | 0.50 | 0.51 | 0.56 | 0.53 |
| distill | **0.54** | **0.63** | **0.60** | 0.55 |
| LERC | 0.53 | 0.56 | 0.54 | **0.60** |
| distill (gold) | **0.54** | 0.61 | 0.59 | 0.58 |
| CFR | 0.53 | 0.62 | 0.59 | **0.60** |

Table 2: In-domain experimental results on AVeriTeC test subset ($n = 200$). Numbers marked with ▪ are statistically significant w.r.t. baseline Contriever at $o = 0.05$ under 10,000 bootstrapped samples. CFR is what we call the model finetuned on *distill (gold) + LERC*.

**HotpotQA** is a question answering dataset featuring multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. The questions require finding and reasoning over multiple supporting documents to answer. There are no claims in this dataset, so we set $c_i = q_i$ and retrieval is done with just the question.

### 4.3 Baselines

We report performance of several widely-used retrievers as baselines: **BM25**, **Contriever** (Izacard et al., 2022) and Contriever fine-tuned on the MS MARCO dataset (Campos et al., 2016) (**Contriever MSM**). We also compare against an additional Contriever baseline. We use **ROUGE-F1** supervision similar to the LERC setup, except long answers were evaluated using ROUGE overlap scores. This tests whether our approaches outperform a simple method for answer matching.

## 5 Results

### 5.1 AVeriTeC

The results for AVeriTeC are shown in Table 2. We find that *distill* performs the best in most metrics but for veracity. The 6% gain in top doc relevance reflect our retriever's ability to correctly identify more relevant documents in our evaluation set.

As expected, we find that using ROUGE as a long answer overlap metric to generate $\{D^+, D^-\}$ works poorly as seen by the ROUGE-F1 baseline.

| Model | ClaimDecomp | | | FEVER | | HotpotQA | |
|---|---|---|---|---|---|---|---|
| | LERC | Top Doc Relv. | Veracity | Top Doc Relv. | Veracity | LERC | Top Doc Relv. |
| BM25 | 0.54 | 0.30 | 0.30 | 0.43 | 0.55 | 0.28 | 0.21 |
| Contriever | 0.64 | 0.32 | 0.32 | 0.49 | 0.58 | 0.33 | 0.27 |
| Contriever MSM | 0.64 | 0.31 | 0.34 | 0.52 | 0.61 | 0.34 | 0.31 |
| gold | 0.64 | 0.30 | 0.28 | 0.48 | 0.56 | 0.32 | 0.30 |
| distill | 0.64 | **0.39** | 0.32 | **0.57** | 0.61 | 0.34 | 0.26 |
| LERC | 0.65 | 0.31 | 0.31 | 0.55 | 0.61 | 0.34 | 0.30 |
| distill (gold) | **0.66** | 0.37 | 0.34 | 0.56 | 0.61 | 0.35 | **0.32** |
| CFR | 0.65 | 0.32 | 0.34 | **0.57** | **0.63** | **0.36** | **0.32** |

Table 3: Out-of-domain experimental results on ClaimDecomp, FEVER, and HotpotQA test subset (n=200 for each dataset). Numbers marked with ▪ are statistically significant w.r.t. baseline Contriever at $p = 0.05$ under 10,000 bootstrapped samples from the respective test subset.

Comparing the average LERC score between baseline Contriever and Contriever finetuned on *LERC*, we find a 5% gain in the average LERC score on the evaluation set. This is also backed by a 6% increase in downstream veracity classification performance, indicating our improved ability to answer questions transfers to actually fact-checking the claim. We also see that the models finetuned with LERC signals (*LERC* and CFR) reflect the strongest improvements in veracity classification. CFR also excels in top doc relevance and other upstream metrics. This indicates evaluating answers derived from documents may help downstream performance on fact-checking more than other supervision signals.

**Lexical overlap** We find that *gold* supervision (using AVeriTeC annotated gold evidence) performs poorly across all metrics. We hypothesize two reasons for this: 1) the evidence lacks significant token overlap with the claim/subquestion and 2) gold annotation involves human reasoning and assumptions which are too complex for the unfinetuned retriever to model in its document embedding space. In fact, the average ROUGE-F1 score between $[c_i; q_{ij}]$ and highest overlapping gold document is only 0.11 compared to 0.25 for the top-ranked document from the wild (see Appendix B.2). This discrepancy comes from examples where the annotated evidence document is based on a related entity not mentioned in the claim or question, which is very challenging to recover without additional context. In other cases, modeling the annotated gold evidence is challenging because it contains new information that is not known from the claim or subquestion alone. Therefore, supervising with only gold documents doesn't effectively help the retriever learn.

### 5.2 Out-of-domain results

Results on out-of-domain datasets are in Table 3.

**ClaimDecomp** We find that our gains translate to ClaimDecomp, with *distill (gold)* demonstrating significant improvements in both LERC and top doc relevance. Examples in this dataset contains both explicit and implicit subquestions, while AVeriTeC subquestions are mostly explicit. Since we use subquestions for retrieval, improvement in top doc relevance may reflect an ability to surface better documents for ambiguous implicit subquestions, which is something baseline retrievers struggle with. An example of this is seen in Appendix D, where our finetuned retriever model is able to accurately capture the focus on lack of funding presented in the question. Even though baseline Contriever selects a document detailing the Amtrak incident with high lexical overlap with the claim and query, the document itself is not useful for answering the question. Using CFR, we see a 2% increase in downstream veracity classification performance.

**FEVER** We also find that our system gives gains on FEVER compared to BM25, Contriever, and Contriever MSM. Our retriever selects relevant top documents more often and yields improved downstream veracity performance.

**HotpotQA** For HotpotQA, we find that *distill (gold) + LERC* performs the best across LERC and top doc relevance. We notice the strongest gains come from including LERC-based supervision, which indicates our retriever may learn to identify answer documents that contain little overlap with the claim. This is especially useful in

multi-hop settings where the answer document cannot be found in one step from the query.

## 6 Retriever Reasoning Capabilities

Our hypothesis about our contrastive training was that it would impart a greater ability for our retriever to "reason" about content rather than directly locating an answer. We conduct an additional study of whether our retriever can exhibit basic 1-hop reasoning capabilities via a synthetic data experiment. We construct positive and negative documents where the positive documents do not directly state the answer, similar to what we found in several AVeriTeC examples.

### 6.1 Synthetic Data Generation

We build these examples by few-shot prompting GPT-4 with synthetic documents written by humans. Our data generation approach takes as input a claim/question pair $(c_i, q_{ij})$ from AVeriTeC and produces a document set $\{d^+, d^-, d_1^-, d_2^-, d_3^-, d_4^-\}$. We generate data for $(c_i, q_{ij})$ pairs from the validation set described in Section 4. The positive document $d^+$ is the only document that contains an answer to the question. Document $d^-$ is a "hard negative" document, which is a document that appears *highly* relevant to the query $[c_i; q_{ij}]$ but does not contain an answer. The 4 other documents $d_1^-, \ldots, d_4^-$ are additional negative documents built from alternate subquestions about the claim.

The **positive document** is a paragraph that supports an answer to the question, but only indirectly. When prompting (Appendix E.2), we require that a clear reasoning hop must be made to recover an answer from the positive document. Therefore, a retrieval system that simply looks for query-document token overlap may not be able to find such documents because the answer is usually not presented in terms of the question.

The **hard negative document** is a paragraph that looks highly relevant to the claim/question, but doesn't actually support an answer. In the prompt, we specify that the document should appear relevant but not support an answer, and further enforce this with few-shot examples (see Appendix E.2). In Appendix F.2, the hard negative document correctly discusses the federal judges Trump nominated. However, it does not contain any information about *how many* judges he nominated, deeming it useless for answering the question about the claim.

| Model | MRR |
|---|---|
| BM25 | 0.49 |
| Contriever | 0.68 |
| Contriever MSM | 0.75 |
| gold | 0.72 |
| distill | **0.80** |
| LERC | 0.72 |
| distill (gold) | **0.80** |
| CFR | 0.79 |

Table 4: Results for 200 examples of synthetically generated data. Numbers marked with ■ are statistically significant w.r.t. baseline Contriever at $p = 0.10$ under 10,000 bootstrapped samples from the respective test set.

The **remaining negative documents** are built by generating alternate subquestions similar to $q_{ij}$ but without overlapping answers. Then, we generate documents that contain answers to these distractor subquestions. An example can be found in Appendix F.1 along with the prompt in Appendix E.2.

### 6.2 Results

We evaluate our retrievers on their ability to score the positive document closer to the query than the negative distractor documents. We measure this via MRR of the positive document across ranking the six documents (positive, hard negative, and 4 alternate question negatives). The results are displayed in Table 4. We find a statistically significant gain in our finetuned model's ability to surface the positive document over other distractor documents. CFR achieves an MRR of 0.79 compared to baseline Contriever (0.68). This supports our hypothesis that finetuning on our supervision signals improves the ability of the retrieval model to find information only indirectly related to the claim.

## 7 Conclusion

This work presents an improved retrieval system, CFR, for fact-checking complex claims. We present two supervision signals for finetuning retrievers under a contrastive objective, and their integration results in improved downstream veracity classification. Furthermore, CFR is able to improve retrieval in settings where inferences are required to identify the correct documents. The gains found in this paper encourage explorations into improving retrieval for fact-checking, as surfacing relevant information proved to be a hard task even for SOTA dense retrievers.

## Limitations

There are a few limitations of our current approach. First, using LERC as an answer equivalence metric requires us to shorten both the gold and candidate answer. The answer compression step loses information that may play a role in verifying hard examples. Therefore, developing a good long answer equivalence metric can help build an even better retrieval system for fact-checking. Such equivalence metrics can also be useful for evaluation: the long-form explanation of why a claim is true or false may be more important than the veracity judgment itself, but this is difficult to assess in an automated way.

Second, this work focuses on the second-stage retrieval step. Building optimized queries for first stage retrieval may yield a better document corpus for second stage, especially for hard examples where little information has been published. However, indexing the necessary documents for the broad set of claims we use involves web-scale indexing, which is beyond the scope of this project.

Finally, this work considered English-language political claims. We note that claims in multimedia (e.g., in memes or videos), claims in other languages, and claims in specialized domains such as COVID-19 misinformation may present distinct challenges. However, we believe that our framework is flexible enough for future work to be able to build on it and train retrievers for these settings as well.

## Ethical Considerations and Risks

This paper presents a retrieval method that seeks to advance the state of the art in automated fact-checking. However, despite recent progress in this area and systems that combine retrieval systems like ours with LLMs (Schlichtkrull et al., 2023; Chen et al., 2024), we stress that these system are not yet ready for deployment. We believe these systems have use to aid professional fact-checkers in their work, since enabling them to quickly find information can aid them to more rapidly check claims. However, these systems cannot produce reliable fact-checks without a human in the loop, as demonstrated by the veracity numbers in this work. Moreover, there is not necessarily a single objective truth about every claim, and a judgment may depend on the reliability of primary sources and other factors which are beyond the scope of this work.


## Acknowledgments

This work was partially supported by Good Systems,[2] a UT Austin Grand Challenge to develop responsible AI technologies, NSF CAREER Award IIS-2145280, and the NSF AI Institute for Foundations of Machine Learning (IFML). We thank the UT Austin NLP community for revisions and feedback on earlier drafts of this paper.


## References


Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. 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. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv*, abs/1611.09268.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. Mocha: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings*


[2] https://goodsystems.utexas.edu/

*of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv*, abs/2312.10997.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *arXiv eprint 2112.09118*.

Gautier Izacard and Edouard Grave. 2022. Distilling knowledge from reader to retriever for question answering. *arXiv eprint 2012.04584*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the Preference Gap between Retrievers and LLMs. *ArXiv*, abs/2401.06954.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv*, abs/2203.05115.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Artsiom Sauchuk, James Thorne, Alon Y. Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeritec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. *ArXiv*, abs/2301.12652.

Prakhar Singh, Anubrata Das, Junyi Jessy Li, and Matthew Lease. 2022. The case for claim difficulty assessment in automatic fact checking. *arXiv eprint 2109.09689*.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

# A Implementation Details

## A.1 Computational Details

The finetuned models were BERT base uncased (110M parameters). Hyperparameter optimization was done via grid search on the learning rate and batch size. For learning rate, we searched $\{1e-5, 2e-5, 4-e5\}$. For batch size, we searched $\{4, 8, 16, 32, 64\}$.

- Infrastructure: 2 NVIDIA Quadro RTX 8000
- GPU Hours (training): approx. 3 hours
- GPU Hours (eval): approx. 1 hour
- Epochs: 12
- Best Learning Rate: 2e-5
- Best Batch Size: 32

## A.2 Experimental Setup

Besides chunking into 200 token spans, document text is not further preprocessed. During training, data was mapped into tuples of the form containing one positive and negative $(c_i, q_{ij}, d^+, d^-)$. That is, if a claim/question pair contains 2 positive and 3 negative paragraphs, it becomes $2 \cdot 3 = 6$ separate data points. These were then shuffled and batched to be fed to the retriever. In contrastive training we use in-batch negatives.

## A.3 Parameters for Packages

- Used rouge-score (v0.1.2) to compute ROUGE-F1 scores. Used `rougeL` (longest common subsequence) with stemming set to True.
- Used openai (v1.34.0) for GPT-4 chat completion. Set temperature setting to 0.2.

## A.4 Scientific Artifacts

- **AVeriTeC** [License] Free to copy, redistribute, and build upon this material given citations and a link to the license. AVeriTeC contains English-language real-world claims mainly in politics gathered from 50 different fact-checking organizations.
- **FEVER** [License] Data annotations incorporate material from Wikipedia, which is licensed pursuant to the Wikipedia Copyright Policy

- **HotpotQA** [License] Free to copy, redistribute, and build upon this material given citations and a link to the license
- **Contriever** [License] Free to copy, redistribute, and build upon this material given citations and a link to the license

# B ROUGE-based Methods

## B.1 ROUGE-based Answer Matching

ROUGE overlap between long answers works is a poor supervision signal because answer strings are typically quite complex. Table 5 illustrates this: although both long answers are conveying the same fact that Nigeria experienced 29 years of military rule, extra details or differences in phrasing can lead to low ROUGE scores despite the answers being semantically equivalent. The opposite may also occur: long answers which contain high lexical overlap may be topically similar but completely different in their key points, creating a false positive example. We also investigated semantic similarity measures like BERT score to assess answer equivalence. Compared to short answer LERC, BERT score tended to work poorly for complex long answers as seen in AVeriTeC. By contrast, using a short answer extraction yields a perfect signal in this case.

## B.2 ROUGE-based Token Overlap

See Table 6. The token overlap between the retriever query (claim+question) and the AVeriTeC annotated gold document is only 0.11, whereas with the top retrieved document it is 0.25. This means using tokens in the query to surface the gold document is not easy.

# C LERC Experiments

## C.1 LERC Quality Check

We evaluate the selection of $\{D_l^+, D_l^-\}$ by manually annotating 10 examples. The task was to select the positive context document given a shuffled, unlabeled $\{D_l^+, D_l^-\}$. We selected the positive document correctly in 60% of examples. Note the positive document here is the one with the highest LERC score (i.e., contains an answer which most closely matches the gold answer). However, the two human annotators agreed on 90% of examples. By investigating the failure cases, we found that LERC-based metrics are sensitive to selecting false negative documents, as human agreement indicated a negative document was more "relevant" to the claim/question than the labeled positive document 40% of the time. Oftentimes, the misclassified document contained a reasonable answer to the question but mismatched the gold answer (hence explaining the low LERC score). This revealed that while LERC can identify strong positive documents, it comes with the risk of including relevant documents as negative contexts.

## C.2 LERC-Human Agreement

In another preliminary study, we manually annotated 22 examples with a fine-grained score from 0-1 reflecting how closely we think the shortened candidate answer matches the shortened gold answer. Across three annotators, we found Kendall's tau agreement scores of 0.55, 0.49, and 0.55 with LERC (Table 7). This indicated human judgments of short answer equivalence correlate well with LERC, making it a viable answer equivalence metric to use as supervision.

# D ClaimDecomp Example

See Table 10

# E GPT-4 Prompts

### E.1 Relevance Prompt

You will be given a claim, a question about the claim, and a passage. Your job is to check whether the passage contains information that supports an answer to the question. You will only output "Yes" or "No".

**Claim:** Hunter Biden had no experience in Ukraine or in the energy sector when he joined the board of Burisma.

**Question:** Did Hunter Biden have any experience in the energy sector at the time he joined the board of the Burisma energy company in 2014?

**Passage:** Hunter Biden , Burisma , Ukraine , and Joe Biden explained - Vox And during the bulk of this troubled period in Hunter ' s life , he was fortuitously on the board of a Ukrainian energy company...

### E.2 Synthetic Data Generation Prompt

You will be provided with a claim and a question about the claim. Your job is to generate two evidence paragraphs:
**(1) Positive:** A paragraph that supports an indirect answer to the claim. It requires a reasoning hop to arrive at the answer. You can make up the answer to the question, but it should only come with a reasoning step.
**(2) Hard Negative:** A paragraph that looks highly relevant to the claim/question, but doesn't actually support an answer Neither paragraph can use "claim" or "question" - they must stand alone and mimic the style of real evidence documents found on the web.

| | Gold Answer | GPT-4 Answer | Score |
|---|---|---|---|
| Long Answer + ROUGE-F1 | Nigeria returned to democracy in 1999, after two long periods of military rule—1966–79 and 1983–98—during which the military wielded executive, legislative, and judicial power | Nigeria experienced military rule for a total of 29 years after independence: from 1966 to 1979 and from 1983 to 1998. | 0.22 |
| Short Answer + LERC | 29 years | 29 years | 1 |

Table 5: Comparison of long answer ROUGE and short answer LERC. The two long answers are effectively conveying the same thing, but the ROUGE-F1 score is only 0.22. However, answer shortening + LERC yields a perfect equivalence score of 1.

| | gold | top_doc |
|---|---|---|
| ROUGE-F1 | 0.11 | 0.25 |

Table 6: Comparing token overlap across 200 examples between $[c_i; q_{ij}]$ and the best annotated gold document or the top-ranked document from the wild (retriever is baseline Contriever).

| Annotators | Kendall $\tau$ |
|---|---|
| 1 / LERC | 0.55 |
| 2 / LERC | 0.49 |
| 3 / LERC | 0.55 |
| 1 / 2 | 0.38 |
| 2 / 3 | 0.40 |
| 1 / 3 | 0.40 |

Table 7: Inter-annotator agreement across 20 examples and 3 annotators. 2/3 refers to the agreement between annotators 2 and 3

Here are some examples:
**Claim:** Former President Donald Trump who lost the popular vote by 3 million has nominated a full third of The United Supreme Court, as of 13th October 2020.
**Question:** How many federal judges did Trump nominate?
**Positive:** Two weeks ago in October Trump nominated multiple members of the Supreme Court. He started by nominating John Jacobs and Patricia McConnell, both of whom have supported Republican policies for many years. He made these judicial appointments despite mass disagreement, highlighting his goal to secure conservative ideals in the judiciary. Last week, he also appointed Max Dermott, making him the third Supreme Court justice nominated by Trump.
**Hard Negative:** Former President Trump nominated highly conservative Supreme Court justices back in October of 2020. His appointments were largely composed of conservative Republicans with long standing connections to Trump. He made these appointments in accordance with mass public support.
**Explanation:** The reasoning step in the positive parargaph is to realize "third of the Supreme court" means 3 out of 9 judges. The positive paragraph correctly lists 3 judges (John Jacobs, Patricia McConnell, and Max Dermott). The hard negative paragraph discusses his appointments but offers no information on how many judges he appointed.

Here is another example:
**Claim:** Anthony Fauci the NIAID director is a democrat.
**Question:** Is Anthony Fauci the NIAID director registered with a political party?
**Positive:** Two weeks ago, a new rule was passed in the NIAID which bans any director from holding political affiliations. In fact, it's even stricter than this - the same rule states no NIAID director is allowed to even register with a political party or participate in elections.
**Hard Negative:** Anthony Fauci has maintained a long standing relationship with Democratic presidential nominee Jacob Wallace. They were childhood friends who grew up together, and Fauci has also openly supported some of Wallace's policies. However, Fauci is historically known to stray away from politics and media.
**Explanation:** The reasoning step in the positive paragraph is to realize NIAID directors cannot register to political parties. Anthony Fauci is an NIAID director according to the claim, therefore he cannot be registered with a political party. The hard negative paragraph mentions his friendship with a Democratic presidential nominee, but this does not imply he is a registered Democrat.

Here is one final, slightly harder example:
**Claim:** Robert E. Lee, commander of the Confederate States Army during the American Civil War, was not a slave owner.
**Question:** Was Robert E. Lee a slave owner?
**Positive:** Many commanders during the Civil War era managed and inherited slaves through their family estates. Robert E. Lee was the commander for the Confederate States Army during the Civil War, and the Confederate states were in support of slavery.
**Hard Negative:** Commander Robert E. Lee led the Confederate States Army during the American Civil War. In the South, many slaves were forced to fight in the army under Robert E. Lee against the Union states. Slaves as soldiers were given poor equipment and placed on the front lines of defense.
**Explanation:** The reasoning step in the positive paragraph is to realize many commanders inherited slaves, and Robert E. Lee was a commander. Therefore it is likely that he might have also had slaves. The hard negative paragraph discusses the role of slaves in the war, but doesn't contain information on whether Robert E. Lee personally owning slaves. Notice even the positive paragraph doesn't contain a direct answer, but it is still

more relevant to the question than the hard negative.

Now, please generate a positive and hard negative paragraph with an explanation for the following claim/question pair:

**Claim:** Hunter Biden had no experience in Ukraine or in the energy sector when he joined the board of Burisma.

**Question:** Did Hunter Biden have any experience in the energy sector at the time he joined the board of the Burisma energy company in 2014?

### E.3 QA Prompt

As a professional fact-checker, your task is to ONLY use the passage to answer the following question about the claim. Keep your answer short (only 1-2 sentences)

**Passage:** Hunter Biden , Burisma , Ukraine , and Joe Biden explained - Vox And during the bulk of this troubled period in Hunter ' s life , he was fortuitously on the board of a Ukrainian energy company...

**Claim:** Hunter Biden had no experience in Ukraine or in the energy sector when he joined the board of Burisma.

**Question:** Did Hunter Biden have any experience in the energy sector at the time he joined the board of the Burisma energy company in 2014?

### E.4 FEVER Veracity Prompt

As a professional fact-checker, your task is to use the following claim and evidence document to determine the veracity of the claim. You must ONLY respond with either SUPPORTS, REFUTES, or NOT ENOUGH INFO

**Claim:** Great white sharks do not prefer dolphins as prey.

**Passage:** Do Sharks Eat Dolphins ? [ Explained ] - Ocean Fauna Did you know that sharks are often considered the ocean ' s top predators ? Well , here ' s an interesting twist : killer whales , which are actually a type of dolphin , are the ultimate predators that can effortlessly take down a shark . But what about other dolphin species ? Do sharks eat dolphins ? Not all sharks eat dolphins , but some species do feed on them . Great whites , tiger sharks , and bull sharks are among the ones that go for it . In this article , I will discuss the types of dolphins that sharks typically consume and how they do it . Are Dolphins Prey Items for Sharks ? Yes , dolphins can indeed be prey items for certain species of sharks . While not all sharks view dolphins as a potential meal , some of the larger and more predatory species , such as Great whites , tiger sharks , and bull sharks , have been known to feed on them . Young and sick dolphins are particularly vulnerable to shark predation , as they have less mobility and are easier to catch . Sharks typically hunt

Now please respond with one of SUPPORTS, REFUTES, or NOT ENOUGH INFO

## F Synthetic Data Examples

### F.1 Generated Example

See Table 8.

### F.2 Human Written Example

See Table 9.

| | |
|---|---|
| Claim | The Obama-Biden administration secretly launched a surveillance operation on the Trump campaign. |
| Question | Was there a surveillance operation on the Trump campaign? |
| Positive | In 2016, it was revealed that the FBI had launched an investigation into potential collusion between the Trump campaign and Russian operatives. This investigation, known as "Crossfire Hurricane," involved surveillance measures such as wiretaps and informants to gather intelligence. Although the operation was conducted by the FBI, it was during the Obama-Biden administration, raising questions about the extent of their involvement and oversight in the surveillance activities. |
| Hard Negative | During the 2016 presidential campaign, the Trump campaign faced numerous allegations of collusion with foreign entities. These allegations led to a series of public investigations and hearings, where various campaign officials were called to testify. The media extensively covered these events, and the public was made aware of the ongoing scrutiny faced by the Trump campaign. However, no direct evidence of a secret surveillance operation by the Obama-Biden administration was presented during these hearings. |
| Alt Q1 | What were the objectives or reasons behind the alleged surveillance operation on the Trump campaign? |
| Alt Q1 Negative ($d_1^-$) | The alleged surveillance operation on the Trump campaign, purportedly initiated by the Obama-Biden administration, was reportedly driven by concerns over potential foreign interference in the 2016 U.S. presidential election. Intelligence agencies had gathered evidence suggesting that Russian operatives were attempting to influence the election's outcome, raising alarms about possible collusion between the Trump campaign and Russian entities. The objective was to safeguard the integrity of the electoral process and ensure that no foreign power could unduly sway the democratic proceedings. Additionally, the surveillance aimed to uncover any illicit activities that could compromise national security. These actions were framed within the broader context of protecting American democratic institutions from external threats. |
| Explanation | The positive paragraph mentions the FBI's "Crossfire Hurricane" investigation, which involved surveillance measures and occurred during the Obama-Biden administration. This requires the reader to make the reasoning hop that the administration might have had some level of involvement or oversight. The hard negative paragraph discusses public investigations and hearings related to the Trump campaign but does not address the existence of a secret surveillance operation by the Obama-Biden administration. |

Table 8: Example of a synthetic example generated from our procedure. The explanation indicates the reasoning hop required to surface the positive paragraph, as well as the complexity of the hard negative.

| | |
|---|---|
| Claim | Former President Donald Trump who lost the popular vote by 3 million has nominated a full third of The United Supreme Court, as of 13th October 2020. |
| Question | How many federal judges did Trump nominate? |
| Positive | Two weeks ago in October Trump nominated multiple members of the Supreme Court. He started by nominating John Jacobs and Patricia McConnell, both of whom have supported Republican policies for many years. He made these judicial appointments despite mass disagreement, highlighting his goal to secure conservative ideals in the judiciary. Last week, he also appointed Max Dermott, making him the third Supreme Court justice nominated by Trump. |
| Hard Negative | Former President Trump nominated highly conservative Supreme Court justices back in October of 2020. His appointments were largely composed of conservative Republicans with long standing connections to Trump. He made these appointments in accordance with mass public support. |
| Explanation | The reasoning step in the positive parargaph is to realize "third of the Supreme court" means 3 out of 9 judges. The positive paragraph lists 3 judges (John Jacobs, Patricia McConnell, and Max Dermott). The hard negative paragraph discusses his appointments but offers no information on how many judges he appointed, which is what the question is asking. |

Table 9: Example of a human annotated positive and hard negative example.

**Claim:** Charles Schumer stated on May 13, 2015 in remarks to reporters: "It is simply a fact that insufficient funding for Amtrak has delayed the installation" of a positive train control safety system. "To deny a connection between the (derailment in Philadelphia) and underfunding Amtrak is to deny reality."
**Question:** Is there a connection between the derailment in Phil. and underfunding Amtrak?

| Contriever | distill (gold) + LERC |
|---|---|
| Latest safety technology wasn ' t fully installed at site of deadly Amtrak derailment south of Seattle - CBS News approached sharp curves at more than double the speed limit . A Metro-North train crashed in New York City in 2013 , killing four people , when an engineer with sleep apnea dozed off . An Amtrak train crashed in Philadelphia in 2015 , killing eight people , when investigators say the engineer was distracted by radio traffic and lost his bearings . Positive train control was installed on 23 percent of the nation 's passenger route miles and 37 percent of freight route miles as of July , the last time the Federal Railroad Administration updated its online tracker for the technology . It is activated on the tracks Amtrak owns along the Northeast Corridor , from Boston to Washington , D.C. , and on Amtrak 's Michigan line . Many of its locomotives are equipped for positive train control . Throughout the rest of the country , Amtrak operates on track owned by freight carriers and other entities that have made varying progress on installing the technology . The new $ 180.7 million route was designed to speed up service by removing passenger trains from a route along Puget Sound that 's bogged down by curves , single-track | Explainer : Positive Train Control and the Amtrak 188 derailment - WHYY Positive Train Control ( PTC ) would have prevented Amtrak 188 from derailing Tuesday , National Transportation Safety Board lead investigator Robert Sumwalt said this week . Amtrak was intending to install the safety system on the Northeast Corridor by the end of year , pursuant to an unfunded congressional mandate under the Rail Safety Improvement Act . **Amtrak CEO Joseph Boardman has said that , with more funding , Amtrak could have implemented PTC sooner** . The stretch of the Northeast Corridor where the derailment occurred currently uses an older system , Automatic Train Control ( ATC ) . On the southbound route , the ATC " enforces " — automatically stops — a train if it is travelling above 45 miles per hour . The northbound side , where Amtrak 188 was travelling , does not . PTC is essentially a smarter version of ATC . Whereas ATC relies on the signal and fixed block system trains have operated on for decades , PTC uses a GPS and radio technology to locate where the trains are along the track . ATC only knows when a train trips a signal wire entering into another large stretch between interlockings or |
| **Answer from GPT:** The passage does not provide information on Amtrak's funding levels or directly link underfunding to the derailment in Philadelphia. | **Answer from GPT:** Yes, according to Amtrak CEO Joseph Boardman, more funding could have allowed Amtrak to implement PTC sooner, which would have prevented the derailment |

Table 10: Comparison of top-1 document on an example from ClaimDecomp between unfinetuned Contriever (left) and CFR model (right). The finetuned retriever is able to surface a document about funding, which is the key aspect the question is targeting.