

The Automated Verification of Textual Claims (AVeriTeC) Shared Task

Michael Schlichtkrull^{1,2}, Yulong Chen², Chenxi Whitehouse^{2,5}, Zhenyun Deng²,
Mubashara Akhtar⁴, Rami Aly², Zhijiang Guo², Christos Christodoulopoulos³,
Oana Cocarascu⁴, Arpit Mittal⁵, James Thorne⁶, Andreas Vlachos²

¹Queen Mary University of London, ²University of Cambridge,

³Amazon AGI, ⁴King’s College London, ⁵Meta, ⁶KAIST

m.schlichtkrull@qmul.ac.uk, {yc632,cj507,zd302,rmya2,zg283,av308}@cam.ac.uk

chrchrs@amazon.co.uk, {mubashara.akhtar,oana.cocarascu}@kcl.ac.uk

thorne@kaist.ac.kr, arpitmittal@meta.com

Abstract

The Automated Verification of Textual Claims (AVERITEC) shared task asks participants to retrieve evidence and predict veracity for real-world claims checked by fact-checkers. Evidence can be found either via a search engine, or via a knowledge store provided by the organisers. Submissions are evaluated using the AVERITEC score, which considers a claim to be accurately verified if and only if both the verdict is correct and retrieved evidence is considered to meet a certain quality threshold. The shared task received 21 submissions, 18 of which surpassed our baseline. The winning team was TUDA_MAI with an AVERITEC score of 63%. In this paper we describe the shared task, present the full results, and highlight key takeaways from the shared task.

1 Introduction

Automated fact-checking (AFC) has been proposed as an assistive tool for beleaguered fact-checkers (Cohen et al., 2011; Vlachos and Riedel, 2014), whose work is crucial for limiting misinformation (Lewandowsky et al., 2020). This has inspired applications in journalism (Miranda et al., 2019; Dudfield, 2020; Nakov et al., 2021) and other domains, e.g. science (Wadden et al., 2020). Substantial progress has been made on common benchmarks, such as FEVER (Thorne et al., 2018a) and MultiFC (Augenstein et al., 2019). Nevertheless, existing resources have recently come under criticism. Many datasets (for example, Thorne et al. (2018a); Schuster et al. (2021); Aly et al. (2021)) contain purpose-made claims derived e.g. from Wikipedia, and are thus not representative of real-world use cases. Datasets that *do* contain real-world claims either lack evidence annotation (Wang, 2017), or suffer issues resulting from superficial automated evidence annotation (Glockner et al., 2022).

Claim: *The USA has succeeded in reducing greenhouse emissions in previous years.*

Date: 2020.11.2 **Speaker:** Morgan Griffith

Q1: What were the total gross U.S. greenhouse gas emissions in 2007?

A1: In 2007, total gross U.S. greenhouse gas emissions were 7,371 MMT.

Q2: When did greenhouse gas emissions drop in US?

A2: In 2017, total gross U.S. greenhouse gas emissions were 6,472.3 MMT, or million metric tons, carbon dioxide.

Q3: Did the total gross U.S. greenhouse gas emissions rise after 2017?

A3: Yes. After 3 years of decline, US CO2 emissions rose sharply last year. Based on preliminary power generation, natural gas, and oil consumption data, we estimate emissions increased by 3.4% in 2018.

Verdict: Conflicting Evidence/Cherry picking.

Figure 1: Example instance from AVERITEC. Given a claim and associated metadata, participating systems must first retrieve appropriate evidence. Then, they must output a verdict for the claim given that evidence.

The AVERITEC dataset was constructed to overcome these limitations (Schlichtkrull et al., 2023a). AVERITEC combines real-world claims with evidence from the web. The process of evidence retrieval is broken down into question generation and answering, providing a structured representation of the evidential reasoning process. The annotation process for AVERITEC was designed to ensure (1) that claims are understandable independently of the fact-checking articles they were sourced from, (2) that the evidence given is sufficient to support the verdicts, and (3) that all evidence used would have been available on the web before the claim was made. This avoids common problems found in previous datasets (Ousidhoum et al., 2022; Glockner et al., 2022).

AVERITEC consists originally of 4,568 examples, collected from 50 fact-checking organizations using the Google FactCheck Claim Search API¹; itself based on ClaimReview². To ensure that systems are evaluated on unseen data, we expanded the (hidden) test set with a further 1,215 claims for the shared task, bringing the total dataset size to 5,783. We furthermore released a “knowledge store” containing, for each claim in the training, development, and test splits, documents which can be used as evidence for that claim. This was done to prevent participants from being limited by the prohibitive cost of the search API we used for evidence retrieval in the original paper (Schlichtkrull et al., 2023a). We also developed an updated version of the baseline for the shared task, which uses the knowledge store. Participants in the shared task were allowed to use evidence from the knowledge store, use a search engine on their own, or combine the two options. Our dataset and baseline are available under a CC-BY-NC-4.0 license at <https://fever.ai/dataset/averitec.html>.

This paper presents a description of the task and dataset, the final test phase leaderboard. We also summarise the submitted system description papers, drawing out commonalities, differences, and lessons. We furthermore carry out additional analysis of the shared task results, including human evaluation. Finally, we reflect on the task, deriving lessons for future work – and further shared tasks – on automated fact-checking. The shared task received 21 submissions. The winning team, TUDA_MAI, achieved a score of 63%, a very significant improvement on the 11% achieved by the baseline system. Nevertheless, there are still plenty of opportunities for further improvement. During the process, we identified an issue with the evidence set provided for participants, which for some claims in the second half of the dataset contained fact-checking articles written by humans about those claims. We release an updated knowledge store at <https://fever.ai/dataset/averitec.html>, where these articles have been removed. We leave open an evaluation page corresponding to the *new* knowledge store³ so that future work can build upon the advances made in this shared task.

¹<https://toolbox.google.com/factcheck/apis>, available under a CC-BY-4.0 license.

²<https://www.claimreviewproject.com/>

³Also available at <https://fever.ai/dataset/averitec.html>

2 Task Description

Participants are given claims and associated metadata, such as the publication date (see Figure 1). Based on this, they must retrieve *evidence* for or against the claims. In the gold annotation, this evidence is broken down into question-answer pairs, naturally enabling multi-hop reasoning. We do not restrict participants to providing evidence in this format, although given the METEOR-based evaluation setup most participants found it beneficial to follow it. When submitting test set predictions, we also required participants to include a URL to an external website for each piece of evidence, corresponding to a webpage providing *backing*. Finally, based on the evidence, participants must predict whether a veracity label from the set *supported*, *refuted*, *not enough evidence*, or *conflicting evidence/cherry-picking*. Unlike the original AVERTIC dataset, we did not require participants to submit a justification for the verdict.

2.1 Dataset

Participants are asked to use the public AVERTIC data for training and validating their systems. To ensure a fairer and more robust evaluation, we constructed a new test set consisting of 1,215 claims, which temporally succeed the original claims, in addition to the original 1000 hidden test set claims of AVERTIC. Like the original test set, these will remain hidden so as to enable future work on the dataset.

Annotation of New Test Set We first collect 2,000 real-world fact-checking articles online from ClaimReview, same source as AVERTIC. Then, we follow the same 5-phase annotation guideline of Schlichtkrull et al. (2023a).

First, given a fact-checking article, an annotator identifies its main claim, collects metadata about it and normalizes the claim by enriching it with necessary context, making it context-independent. Second, given the normalized claim, another annotator generates questions and answers (QAs) with the help of the fact-checking article and the web, and gives a verdict label for the claim. Third, given only the QAs as evidence, a different annotator selects a verdict label for the claim and provides a justification for their choice. At this point, we compare the verdict labels annotated by different annotators. If the labels match, we consider the evidence is sufficient for predicting the veracity; otherwise, we repeat the last two phases as our

Split	Train	Dev	Test (old)	Test (new)
Claims	3,068	500	1,000	1,215
Question / Claim	2.60	2.57	2.57	2.89
Re-annotated (%)	28.1	24.4	25.1	20.0
End date	25-08-2020	31-10-2020	22-12-2021	13-08-2023
Labels (S/R/C/N)	27.6/56.8/6.4/9.2	24.4/61.0/7.6/7.0	25.5/62.0/6.3/6.2	17.3/66.5/4.1/12.1
Types (PS/NC/EPC/QV/CC)	7.8/33.7/57.8/9.6/11.5	5.8/23.8/61.4/13.8/10.8	7.0/21.9/69.8/7.7/11.9	3.5/24.3/71.9/5.2/16.1
Strategies (WE/NCP/FR/EC/SS)	78.8/30.6/6.6/29.9/3.6	88.6/19.0/7.4/27.4/2.0	88.0/19.2/7.7/29.6/1.8	82.4/22.6/10.0/37.6/4.0

Table 1: Statistics for the new test set. For better comparison, we present the statistics for the original dataset. The Labels (%) are Supported (S), Refuted (R), Conflicting Evidence/Cherry-picking (C), and Not Enough Evidence (N). The Claim Types (%) are Position Statement (PS), Numerical Claim (NC), Event/Property Claim (EPC), Quote Verification (QV), and Causal Claim (CC). The Fact-checker strategies (%) are Written Evidence (WE), Numerical Comparison (NCP), Fact-checker Reference (FR), Expert Consultation (EC) and Satirical Source (SS). Note that we for simplicity omitted very low-frequent fact-checker strategies, e.g., Geo-location (0.3%).

fourth and fifth phases, respectively. If the labels given by the fourth and fifth annotators still do not match, we discard this instance. In this way, we obtain 1,215 new instances. Each is annotated with a normalized claim, meta-data, QA pairs as evidence, a verdict label and a justification for it. For the detailed annotation guidelines and procedures, please refer to Schlichtkrull et al. (2023a).

To ensure high quality, we train our annotators before formal annotation. For each phase, annotators are first asked to annotate 10 instances. We then provide feedback and highlight their most frequent and common mistakes. They are then asked to annotate another 10 instances. We select qualified annotators based on their performance on 3 tasks: (1) claim type and fact-checking strategies over 70%+ F -1 scores; (2) 2+ QA pairs per claim; (3) veracity prediction over 50%+ accuracy. These criteria are based on empirical consideration from the earlier AVERITEC annotation (Schlichtkrull et al., 2023a). Finally, we selected 12 qualified annotators from 34 participants.

Comparison between Original and New Test Sets Table 1 presents the statistics of our new test set in comparison with the original AVERITEC dataset. Our new test set (with claims up to 2023) is temporally further removed from the training set (ending in 2020). As such, there can be a domain shift between new and old data, regarding the fact-checking content. However, the majority (66.5%) of claim labels are *refuted*, which is consistent with previous data. Additionally, the distributions of claim labels, claim types and fact-checking strategies are largely similar in terms of their proportions. The new test set has slightly more questions per claim compared to the original one, indicating that the annotation process was at least as thorough.

2.2 Knowledge Store

As mentioned in Schlichtkrull et al. (2023a), reliance on the Google search API made the original baseline prohibitively expensive. Thus, to mitigate the cost, we released a *knowledge store* along with the shared task. The knowledge store contains a collection of potentially useful evidence documents for each claim, obtained via Google search.

We collected the knowledge store using a process inspired by our original baseline. We extracted a variety of search queries using ChatGPT⁴, based on the claim, gold questions, and gold answers. We further used *distractor queries* created by changing entities, dates, and events in the claim, in order to add plausible – but irrelevant – documents to the knowledge store. All queries can be seen in Appendix A. For each query, we collected every URL returned on the first page of the Google Search API. We used the same temporal restrictions as in Schlichtkrull et al. (2023a), ensuring that the included documents would have been available on the web before the claim was made. We also included the annotator-selected evidence documents selected for each claim. We deduplicated and shuffled the documents corresponding to each claim.

We provided the URL for each document, as well as a text version scraped using *trafilatura* (Barbaresi, 2021). The knowledge store includes text scraped from PDF URLs, a step omitted in Schlichtkrull et al. (2023a). Furthermore, for the train and development splits (but not test), we made available the specific Google search query used for each document, as well as the category (see Table 11). The average claim has 955 associated documents, each of which have on average of 6,095 tokens. The most common URL

⁴We used `gpt-3.5-turbo-0125`.

domains for knowledge store documents are, in order, the National Center for Biotechnology Information (NCBI), Wikipedia, Quora, the New York Times, and CNN.

The knowledge store allowed participants to compete without access to a paid search engine. Further, it allowed inexpensive experimentation with a variety of different retrieval strategies. Our construction process for the knowledge store relies on information not available normally to participants, such as the gold question-answer pairs. We found that these were necessary to ensure that good, relevant evidence was included. At the same time, relying on the knowledge store complicates the finding of alternative evidence paths to the one used by our annotators. Exploring alternative evidence paths was easier for systems which directly integrated their own search engine. As such, there were upsides to both strategies.

2.3 Baseline

Our baseline closely follows the approach described in Schlichtkrull et al. (2023a), with the main difference being that, instead of requiring direct access to the paid Google Search API, we use the aforementioned knowledge store. This adjustment aims to reduce the costs of participating in the Shared Task.

Our baseline consists of the following steps. (1) We parse the scraped text into sentences and rank their similarity to the claim using BM25 (Robertson and Zaragoza, 2009), retaining the top 100 sentences per claim. (2) Questions-answer (QA) pairs are generated for these top 100 sentences using BLOOM,⁵ with the 10 most similar claim-QA pairs from the training set used as in-context examples. (3) The QA pairs are then re-ranked using a pretrained BERT model as described in Schlichtkrull et al. (2023a). (4) Finally, using the top-3 QA pairs as evidence, we predict the veracity label of the claim with another pretrained BERT model, as detailed in Schlichtkrull et al. (2023a).

The baseline results are shown in Table 2. We note that on both the development set, the old test set, and the new test set, the shared task baseline and the baseline from Schlichtkrull et al. (2023a) perform similarly. Further details regarding the implementation, knowledge store, and pretrained BERT models are available at <https://huggingface.co/chenxwh/AVeriTeC>.

⁵We used `bigscience/bloom-7b1`.

2.4 Evaluation

The primary evaluation metric for the shared task is AVERITEC score, discussed in depth in Schlichtkrull et al. (2023a). We first compute results for question generation and question-answer generation using Hungarian METEOR score. That is, we use the Hungarian Algorithm (Kuhn, 1955) to find an optimal matching of generated text to reference text in terms of METEOR score. Formally, let $X : \hat{Y} \times Y \rightarrow \{0, 1\}$ be a boolean function denoting the assignment between the first 10 generated question-answer pairs (or questions only) \hat{Y} and the reference question-answer pairs (or questions only) Y . Then, the Q + A score (or Q only score) u is calculated as:

$$u_f(\hat{Y}, Y) = \frac{1}{|Y|} \max \sum_{\hat{y} \in \hat{Y}} \sum_{y \in Y} f(\hat{y}, y) X(\hat{y}, y) \quad (1)$$

where the pairwise scoring function $f : S \times S \rightarrow \mathbb{R}$ is METEOR score (Banerjee and Lavie, 2005) using the NLTK implementation (Bird et al., 2009).

To compute the AVERITEC score, we applied a cutoff of $u_f(\hat{Y}, Y) \geq 0.25$ to determine whether adequate evidence has been retrieved, using the Q + A Hungarian METEOR score. Any claim for which this score is lower than 0.25 receives an AVERITEC score of 0. For claims where the evidence score is higher than 0.25, the AVERITEC score is defined as the accuracy of the predicted verdict (veracity). As also discussed in Schlichtkrull et al. (2023a), both for Q only, Q+A, and AVERITEC score, if a system provided more than 10 QA pairs, all pairs after the 10th were discarded. We note that QA pairs beyond the 10th can still be input to veracity prediction components, and may as such still be useful to some systems.

3 Results

The overall results for the shared task can be seen in Table 2. Each of the 21 participating teams were invited to submit a paper to be reviewed in the FEVER workshop – detailed descriptions for each system can be found in the corresponding papers. 15 system description papers were submitted to the workshop (with a 16th submitted and withdrawn). We analyse the model components discussed in each paper – see Table 3. Below, we present our general observations on the techniques used by participants, as reported in their respective system description papers.

Rank	Team Name	Q only	Q + A	AVERITeC @ .25
1	TUDA_MAI (Rothermel et al., 2024)	0.45	0.34	0.63
2	HUMANE (Yoon et al., 2024)	0.48	0.35	0.57
3	CTU AIC (Ullrich et al., 2024)	0.46	0.32	0.50
4	Dunamu-ml (Park et al., 2024)	0.49	0.35	0.50
5	Papelo (Malon, 2024)	0.44	0.30	0.48
6	UHH (Sevgili et al., 2024)	0.48	0.32	0.45
7	SynApSe (Churina et al., 2024)	0.41	0.30	0.42
8	arioriAveri (Momii et al., 2024)	0.38	0.29	0.39
9	Data-Wizards (Singhal et al., 2024)	0.35	0.27	0.33
10	MA-Bros-H (Mohammadkhani et al., 2024)	0.38	0.24	0.27
11	mitchelldehaven	0.27	0.23	0.25
12	SK_DU (Malviya and Katsigiannis, 2024)	0.40	0.26	0.22
13	UPS (Omar, 2024)	0.31	0.27	0.21
14	FZI-WIM (Liu et al., 2024b)	0.32	0.21	0.20
15	KnowComp (Liu et al., 2024a)	0.32	0.21	0.18
16	IKR3-UNIMIB (Urbani et al., 2024)	0.32	0.24	0.18
17	ngetach	0.37	0.21	0.14
18	VGyasi	0.38	0.22	0.12
19	<i>Baseline</i>	<i>0.24</i>	<i>0.20</i>	<i>0.11</i>
20	InfinityScalers!	0.26	0.19	0.08
21	AYM	0.13	0.12	0.06
22	Factors	0.20	0.14	0.05

Table 2: Overall results for the AVERITeC shared task. Performance is evaluated on the total of 2214 hidden test set examples. Scores are given in Hungarian METEOR for question-only and question-answer performance, and in AVERITeC-score at evidence cutoff 0.25 for total performance (see Schlichtkrull et al. (2023a)).

Knowledge Source Papelo, SynApSe, and KnowComp relied on the Google Search API as knowledge source, while the remaining systems all used our knowledge store. Participants identified shortcomings in both approaches: the knowledge store is guaranteed to include the gold evidence and can be searched with highly performant embedding methods, whereas the search API allows for more freedom in what information can be retrieved (i.e., if generating questions for a different evidence path than the one our annotators used, the knowledge store may not be able to answer those questions). As evidenced by the strong results of Team Papelo, despite the predominance of systems relying on the knowledge store, the Google Search API (with which the knowledge store itself was built) remained a competitive option (see Table 2).

One issue identified by several participants was the scraper we used for the knowledge store, based on Trafilatura (Barbaresi, 2021). Papelo identified how, in 297 out of 500 development examples, at least one gold document was not correctly scraped. Dunamu-ML similarly discussed how the scraper

did not correctly handle evidence from PDFs and videos. In their submission, Dunamu-ML extended the scraper to extract text and transcripts from PDFs and YouTube videos, and noted that this helped performance. When constructing AVERITeC, our annotators filtered out claims requiring multimodal reasoning; all claims in the dataset are textual and can be verified through exclusively textual evidence. Nevertheless, the helpfulness of video transcripts suggests that multimodal evidence can be useful even for that scenario.

Question Generation & Retrieval Most systems employed an LLM-based question generation strategy. That is, they generated questions or queries, and then retrieved evidence based on those questions. Generating questions, rather than simply searching for the claim, was noted by many top-scoring systems to be essential for good retrieval performance. This supports our hypothesis from Schlichtkrull et al. (2023a) that question generation (or query expansion (Mao et al., 2021)) is a key avenue for further gains in retrieval.

Team Name	Evidence	QG	Retrieval	QA	Veracity
TUDA_MAI	KS	GPT-4o	gte_base_en_v1.5	GPT-4o	GPT-4o
HUMANE	KS	Llama-3-8b	BM25 SFR-embedding-2 Llama-3.1-70b	-	Llama-3.1-70b
CTU AIC	KS	GPT-4o	mxbai-large-v1	GPT-4o	GPT-4o
Dunamu-ML	KS	GPT-4	BM25	GPT-4	GPT-4
Papelo	Google	T5-large GPT-4o	-	GPT-4o	GPT-4o
UHH	KS	GPT-4o-mini	BM25 gte_base_en_v1.5	GPT-4o-mini	Mixtral-8x7B
SynApSe	Google	GPT-4o	all-MiniLM-L6-v2	GPT-4o	GPT-4o GPT-3.5 Mistral-7B
aioriAveri	KS	GPT-4o	stella_en_400M_v5	GPT-4o	GPT-4o
Data-Wizards	KS	Phi-3-medium	stella_en_1.5B_v5	Mixtral-8x22B	Mixtral-8x22B
MA-Bros-H	KS	Llama-3-70B	BM25	Llama-3-70B	Llama-3-70B
SK_DU	KS	GPT-4o	BM25 ms-marco-MiniLM-L-12-v2	-	deberta-v3-base
UPS	KS	T5-large	BM25 BERT	-	BERT
FZI-WIM	KS	Llama-3-70B	ms-marco-MiniLM-L-12-v2	Llama-3-70B bart-large-mnli	Llama-3-70B
KnowComp	Google	Llama-3-8b	-	Llama-3-8b	Llama-3-8b
IKR3-UNIMIB	KS	-	BM25 ColBERT	GPT-3.5	BERT

Table 3: Components used by systems that submitted description papers. Systems are ordered based on AVeriTeC-score (see Table 2). - indicates, respectively, that a system directly used claims and nothing else for search queries, that retrieval was done only through a search API with no reranking, and that the answer used was the entire retrieved passage.

Question generation was typically implemented using large-scale LLMs, such as GPT-4o or Llama-3.1-70b. Some systems based on smaller model – HUMANE with Llama-3-8b, UHH with GPT-4o-mini, Data-Wizards with Phi-3-medium, and Papelo with T5 (for the first question only) – also achieved a high question-only score. This suggests that smaller models can be competitive on search query generation.

Several teams – Papelo, SynApSe, and IKR3 – mentioned that they saw benefits from modeling the retrieval task as multi-hop retrieval. That is, instead of retrieving all documents at once, their systems used multiple rounds of retrieval with each round conditional on previous rounds. The benefits of this strategy were also documented in previous FEVER shared tasks, e.g., Malon (2021). Team Papelo further expanded on this strategy, showing that the use of different components at different retrieval steps – T5 for the first question and GPT-4o for subsequent questions – yielded higher performance than using a single-question generation model.

As can be seen in Table 5, high-performing systems tended to generate and submit a high number of questions. This may be a consequence of our evaluation setup – there is no brevity penalty (other than documents past the 10th being ignored), so submitting more evidence documents means a higher chance of recalling the gold evidence. Several teams also noted that even duplicates of the same question could slightly increase their score.

We tested this, and observed baseline performance increase by 2 points QA score and 0.5 points AVeriTeC score when including two additional duplicates of each question. There are two reasons this might happen. First, some generated QA pairs may be the best match for multiple gold QA pairs (i.e. because they are very long, or because other QA pairs are irrelevant to the claim). Duplicating QA pairs means the generated pair can be matched to multiple gold pairs when computing the Hungarian algorithm, marginally increasing overall performance. Second, Hungarian METEOR is computed by averaging over gold question-answer

Team Name	Text	PDF	Table	Metadata	Audio	Video	Image	Other	1 doc	2 docs	3+ docs
TUDA_MAI	0.34	0.35	0.36	0.31	0.31	0.33	0.32	0.33	0.39	0.35	0.31
HUMANE	0.34	0.36	0.38	0.32	0.34	0.32	0.33	0.38	0.41	0.35	0.31
CTU AIC	0.31	0.33	0.36	0.30	0.26	0.30	0.32	0.35	0.33	0.33	0.29
Dunamu-ml	0.34	0.36	0.39	0.31	0.24	0.33	0.34	0.37	0.40	0.36	0.32
Papelo	0.3	0.31	0.32	0.27	0.22	0.29	0.29	0.3	0.35	0.3	0.27
UHH	0.31	0.34	0.36	0.29	0.23	0.31	0.31	0.37	0.37	0.32	0.28
SynApSe	0.29	0.31	0.32	0.25	0.25	0.28	0.28	0.31	0.38	0.32	0.22
arioriAveri	0.28	0.29	0.32	0.26	0.21	0.27	0.27	0.32	0.34	0.29	0.25
Data-Wizards	0.26	0.26	0.28	0.23	0.17	0.27	0.25	0.27	0.36	0.29	0.19
MA-Bros-H	0.23	0.25	0.28	0.22	0.16	0.23	0.22	0.27	0.3	0.26	0.19
mitchelldehaven	0.22	0.23	0.24	0.18	0.19	0.22	0.2	0.22	0.28	0.23	0.19
SK_DU	0.25	0.26	0.27	0.22	0.17	0.25	0.24	0.27	0.34	0.28	0.18
UPS	0.26	0.29	0.31	0.25	0.23	0.27	0.28	0.31	0.29	0.27	0.25
FZI-WIM	0.2	0.22	0.24	0.18	0.12	0.18	0.19	0.21	0.27	0.22	0.15
KnowComp	0.2	0.22	0.23	0.18	0.05	0.18	0.19	0.22	0.29	0.23	0.14
IKR3-UNIMIB	0.23	0.24	0.26	0.19	0.13	0.23	0.21	0.25	0.31	0.25	0.16
ngetach	0.21	0.22	0.23	0.18	0.15	0.19	0.2	0.23	0.24	0.23	0.18
VGyasi	0.21	0.22	0.24	0.2	0.11	0.22	0.2	0.24	0.27	0.24	0.17
<i>Baseline</i>	<i>0.19</i>	<i>0.2</i>	<i>0.23</i>	<i>0.17</i>	<i>0.14</i>	<i>0.19</i>	<i>0.19</i>	<i>0.21</i>	<i>0.24</i>	<i>0.21</i>	<i>0.14</i>
Factors	0.19	0.19	0.21	0.16	0.21	0.18	0.16	0.17	0.24	0.2	0.15
InfinityScalers!	0.11	0.11	0.1	0.08	0.07	0.11	0.1	0.09	0.22	0.12	0.06
AYM	0.13	0.13	0.13	0.1	0.05	0.12	0.11	0.13	0.26	0.14	0.06
Average	0.25	0.26	0.28	0.22	0.18	0.24	0.24	0.26	0.31	0.26	0.2

Table 4: Retrieval results in terms of Q+A Hungarian METEOR, broken down according to 1) the document type of the gold evidence, and 2) the number of gold evidence QA pairs for the claim. The overall best performance on retrieval was achieved by Dunamu-ML.

pairs. If there are more gold pairs than generated pairs, some gold pairs will be *unmatched*. These will receive a score of 0, as the “matched” evidence is the empty string, dragging down the average. Effectively, systems are heavily penalised for generating too *few* questions, and may benefit slightly from generating too *many*.

For evidence retrieval, vector-based dense retrieval systems (Karpukhin et al., 2020) were common, along with BM25 (Robertson and Zaragoza, 2009). Several teams – HUMANE, UHH, SK_DU – proposed hybrid systems where coarse retrieval with BM25 was followed by reranking with a vector-based approach. For vector-based retrievers, the *gte* (Li et al., 2023; Zhang et al., 2024) family of models were popular, and noted by participants to perform well on the task; this includes Stella⁶, an MRL (Kusupati et al., 2022) approach based on *gte*. Several teams noted that their *gte*- or Stella-based retrievers were chosen as they, at the time of the competition, were top performers on the MTEB (Muennighoff et al., 2023) leaderboard.

⁶https://huggingface.co/dunzhang/stella_en_400M_v5

The overall best performing retrieval system was Dunamu-ML, closely followed by HUMANE. In Table 4, we break down performance on retrieval according to which document type the *gold* evidence originated from. We see that Dunamu-ML do have top performance on PDFs and videos (for which they added a custom scraper), but tie respectively with HUMANE and TUDA_MAI on these categories. On the other hand, Dunamu-ML perform better than other systems on tabular and image evidence, while HUMANE is the top performer on Metadata, Audio, and “Other” evidence (used by participants mostly for social media posts, as well to link to external web tools, such as a calculator in support of numerical reasoning).

In Table 4, we also break down retrieval performance by the number of gold evidence question-answer pairs per claim. HUMANE performs the best on claims with only one gold document, narrowly followed by Dunamu-ML. As the number of claims increases, Dunamu-ML takes the lead. With an average of 2.74 questions per claim in the test set, this may explain why Dunamu-ML achieved the overall highest retrieval performance.

Team name	QV	N	E/P	C	PS	S	R	NEE	CE/C	Avg. # Docs
TUDA_MAI	0.64	0.58	0.64	0.64	0.58	0.64	0.73	0.12	0.19	9.3
HUMANE	0.59	0.57	0.58	0.55	0.46	0.76	0.62	0.01	0.12	10.0
CTU AIC	0.57	0.49	0.51	0.52	0.38	0.58	0.58	0.1	0.01	9.89
Dunamu-ml	0.44	0.49	0.5	0.55	0.4	0.69	0.5	0.31	0.12	12.41
Papelo	0.51	0.38	0.5	0.51	0.45	0.45	0.59	0.0	0.0	9.95
UHH	0.46	0.43	0.46	0.48	0.39	0.47	0.54	0.0	0.0	10.0
SynApSe	0.45	0.39	0.43	0.43	0.36	0.42	0.5	0.02	0.21	4.26
arioriAveri	0.44	0.37	0.39	0.4	0.29	0.45	0.44	0.09	0.06	8.98
Data-Wizards	0.37	0.3	0.34	0.32	0.29	0.44	0.36	0.05	0.04	3.0
MA-Bros-H	0.29	0.3	0.26	0.25	0.19	0.4	0.27	0.08	0.0	3.74
mitchelldehaven	0.24	0.26	0.25	0.25	0.16	0.4	0.25	0.0	0.0	5.0
SK_DU	0.27	0.3	0.21	0.15	0.14	0.36	0.22	0.01	0.11	3.0
UPS	0.29	0.18	0.22	0.2	0.21	0.17	0.24	0.08	0.14	10.0
FZI-WIM	0.21	0.25	0.18	0.16	0.21	0.31	0.18	0.12	0.02	2.52
KnowComp	0.16	0.19	0.19	0.15	0.13	0.27	0.19	0.0	0.01	2.55
IKR3-UNIMIB	0.21	0.22	0.17	0.17	0.15	0.28	0.19	0.01	0.05	3.0
ngetach	0.16	0.13	0.14	0.17	0.09	0.0	0.22	0.0	0.0	4.25
VGYasi	0.16	0.11	0.13	0.11	0.10	0.1	0.12	0.22	0.03	3.46
<i>Baseline</i>	<i>0.14</i>	<i>0.16</i>	<i>0.11</i>	<i>0.10</i>	<i>0.06</i>	<i>0.17</i>	<i>0.12</i>	<i>0.0</i>	<i>0.04</i>	3.0
InfinityScalers!	0.04	0.10	0.09	0.08	0.08	0.24	0.04	0.04	0.10	3.52
AYM	0.07	0.06	0.06	0.03	0.10	0.11	0.05	0.0	0.0	1.0
Factors	0.04	0.05	0.05	0.05	0.04	0.13	0.03	0.04	0.01	1.0
Average	0.31	0.29	0.29	0.29	0.24	0.36	0.32	0.06	0.06	5.63

Table 5: We compute separate results based on claim type (QV = Quote Verification, N = Numerical, E/P = Event/Property, C = Causal, PS = Position Statement). We also compute results separated by gold verdict (S = Supported, R = Refuted, NEE = Not Enough Evidence, CE/C = Conflicting Evidence / Cherrypicking). Finally, we report the average number of evidence documents submitted per claim. We note that if a team submitted more than 10 documents for a claim, only the first 10 were used to compute retrieval scores for evaluation.

Veracity Prediction Veracity prediction was also dominated by LLM-based approaches, including GPT-4o, Llama 3.1, and Mixtral. Teams HUMANE and SynApSe note that some fine-tuning was necessary for good performance on veracity prediction. Various teams saw improvements both from full fine-tuning of all the weights, and from fine-tuning with LORA (Hu et al., 2022). Interestingly, one team – Papelo – chose to prevent their veracity prediction system from predicting Not Enough Evidence and Conflicting Evidence, arguing that their prompting-based model too frequently chose these rarer labels. This may explain why calibration was especially helpful for this task.

We note that top-scoring systems tended to use very large models for veracity prediction, such as GPT-4o, Llama-3.1-70b, or Mixtral-8x7b. The superior reasoning capabilities of these cutting-edge models appear especially critical to this stage of the pipeline, unlike for question generation.

Types & Verdicts In Table 5, we provide a detailed breakdown of the results based on claim type (quote verification, numerical claims, event/property claims, causal claims, position statements) and verdict (supported, refuted, conflicting evidence/cherrypicking, not enough evidence). For each category, we report AVERITEC scores on the corresponding subset of the test set.

Systems performed slightly better on quote verification, slightly worse on position statements, and approximately equally well on other claims. This is interesting, as quote verification and position statements are relatively similar tasks. In the former, systems must verify if a person has uttered a quote verbatim; in the latter, systems must verify if a person or organisation holds a specific position (e.g., supporting a policy), but not necessarily verbatim. Verifying position statements often required abductive reasoning, which LLMs are known to struggle with (Dougrez-Lewis et al., 2024).

Among the top performing systems, performance is frequently lower on numerical statements (along with position statements) compared to other claims. This suggests that the gap is smaller for numerical reasoning than other forms of reasoning. As top performers often use very large LLMs, that is suggestive of the type of reasoning gains accomplished by scaling up these models.

In terms of performance across the different labels, there is significant variation. First, systems often have different calibration to predict supported versus refuted claims. As refuted claims dominate (making up approximately two-thirds of the dataset), this yields a significant advantage for some participants. We note that a common strategy among participants was to ignore the rarer veracity labels – not enough evidence, and conflicting evidence. As mentioned e.g. by team Papelo in their system description paper, large language models tend to overpredict these rarer classes. Nevertheless, many top performers, including the winning system, made significant gains on these classes.

Quality Controls on Test Submissions To ensure the reliability of submitted systems, we conducted quality control on our submissions. Here, *reliability* refers to the evidence (QA pairs) being grounded and supported by their retrieved documents. Typically, participants returned answers generated based on retrieved documents; although some systems generated answers e.g. with an LLM, and subsequently matched the answer to a “backing document”.

We first used an automatic method to evaluate the entailment between the answers and the retrieved documents. Specifically, we applied a DeBERTa-large-based NLI model (He et al., 2020)⁷ on all submissions, taking each answer as hypothesis and its corresponding document as premise. Generally, we find that most teams see a small proportion of entailment labels and a large proportion of neutral labels (80%). This can be because the NLI model cannot perform well on out-of-distribution data in a zero-shot setting, in particular when the retrieved document is much longer than the standard NLI premise (e.g., the average document length in words in TUDA_MAI’s submission is over 4,000, while it is around 50 in ANLI (Mishra et al., 2021)).

⁷<https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli>, which demonstrates the best performance on NLI tasks amongst Hugging Face models.

Therefore, we further investigated submissions via manual evaluation. In particular, we focused on instances which the NLI model identified as either *neutral* or *contradiction*, and on the top-4 performing systems (i.e.: TUDA_MAI, HUMANE, CTU AIC and Dunamu-ml). We randomly selected 20 neutral or contradicting instances from each submission, and then performed human evaluation. Given an instance with its corresponding QA pairs and retrieved documents, we identified whether the answers were entailed by the retrieved documents.

Generally, we found that all systems were mostly reliable, with the evidence they generate being supported by the retrieved documents. All answers from TUDA_MAI were extractive from source documents and thus entailed. The answers from the other three systems were more abstractive. Although the answers can contain some hallucination (e.g., generating answers that contradict the retrieved documents by mistake), our manual evaluation found the answers were mostly (HUMANE: 19/20; CTU AIC: 17/20; Dunamu-ml: 12/20) entailed by their associated documents. Errors were typically due to mistakes by the question-answering components, such as taking a snippet from the associated document out of context. Thus, we conclude that the systems evaluated were reliable and find relevant documents that provide useful information for predicting veracity.

4 Human Evaluation of Evidence

Following the approach taken in the first FEVER shared task (Thorne et al., 2018b), we conducted human evaluation of the evidence retrieved by the systems participating in the shared task, motivated by two concerns. First, the incompleteness of the gold evidence annotation, since it is often the case that adequate evidence to determine the verdict for a claim can be found in multiple webpages, as shown in the inter-annotation agreement study of Schlichtkrull et al. (2023a). Second, the inaccuracies of automatic evaluation metrics of textual evaluation, especially in the case of token-matching metrics such as METEOR (Banerjee and Lavie, 2005) used here, but also of more recent neural ones such as FactScore (Min et al., 2023). Thus we can gain a deeper understanding of the quality of the retrieved evidence, and assess how well the AVERITEC scores assigned to the retrieved evidence aligns with human judgements.

Evaluation Process We conducted human evaluation in collaboration with the participating teams. Sixteen top-performing teams were invited to participate in the evaluation. However, teams Dunamuml, mitchelldehaven, and KnowComp did not take part. Each of the remaining thirteen participating teams manually evaluated thirty evidence samples from other participants. Out of these, five were gold-labeled, which were included to assist in the post-processing of the collected annotations and to assess their quality. The evidence samples were randomly selected and evenly distributed across all submitted systems, representing both high- and low-scoring systems, as shown in Table 5.

Figures in Appendix B depict the evaluation form and the instructions provided to human annotators during evaluation. As a first step, we asked annotators to assess whether “at least some part of the evidence” was “non-empty, understandable, and related to the claim.” If so, it was considered eligible for further rating. In addition to assigning a verdict label, we asked annotators to rate retrieved evidence in comparison to provided reference evidence⁸. Annotators rated the evidence on a scale from 1 to 5 across five dimensions:

- (1) **Coverage:** Measures how much of the reference evidence is covered by the predicted evidence, ensuring that the content, meaning, entities, and other key elements of the reference are fully represented in the retrieved evidence.
- (2) **Coherence:** Captures whether the retrieved evidence is coherent, i.e., if all sentences are connected sensibly and the evidence makes sense as a whole.
- (3) **Repetition:** Evaluates whether the retrieved evidence exhibits repetition of its content.
- (4) **Consistency:** Assesses whether the retrieved evidence is semantically consistent and does not contain conflicting information. Unlike coherence, which focuses on how well the information is structured, consistency evaluates whether the arguments presented in the evidence for or against a claim are sound and aligned.
- (5) **Relevance:** Measures how relevant the retrieved evidence is to the content of the claim.

Insights Gained The annotation process resulted in a total of 389 annotations. After filtering out evidence samples that were labeled by evaluators as entirely empty (1%), not understandable (1.8%), or

⁸We provide the exact instruction for rating each criteria in the appendix.

Label/Pred	CE/C	NEE	Refuted	Supported
CE/C	35.7	3.6	53.6	7.1
NEE	5.9	22.1	60.3	11.8
Refuted	3.9	4.9	85.4	5.8
Supported	7.6	0	16.5	76.0

Table 6: Overview of verdict **labelled** by human evaluators (rows) versus system **predictions** (columns).

completely irrelevant to the given claim (9.4%), we were left with 344 valid annotations. Among these, 66 annotations corresponded to gold-labeled samples. Excluding the gold-labeled samples, resulted in a final set of 278 evidence annotations.

Before labeling the system-retrieved evidence, participants were first asked to label the verdict of the retrieved evidence. Table 6 provides an overview of the matching between system-predicted labels (columns) and human-labeled verdicts (rows). While human annotators generally agreed with evidence labeled as refuted or supported, there was less overlap for evidence labeled as NEE and CE/C by the submitted systems.

Analyzing human judgments across the five evaluated dimensions (see Table 10), we find that the majority of predicted evidence was labeled as very coherent, consistent, relevant, and containing limited repetition. However, in the dimension of semantic coverage, approximately 15% of the evidence received a rating of 0, indicating that “the predicted evidence covers none of the reference evidence.” Additionally, around 20% received a rating of 1, meaning that “very little of the reference evidence is covered.” This does not necessarily mean that the evidence is false – low coverage can also occur if the retrieved evidence uses different information, arguments, or sources than the reference evidence. Ideally, we aim for an evidence evaluation that can fairly assess evidence even when it differs from the reference and has low coverage.

To assess the relationship between human scoring and the Hungarian METEOR (see Sec 2.4), we computed both the Spearman correlation coefficient (ρ (Spearman, 1987)) and the Pearson correlation coefficient (r (Pearson, 1896)) as shown in Table 8. Correlations were calculated using both the entire evidence text and the question text only. In both cases, we observed a low correlation between the Hungarian Meteor and the assessed dimensions, with the highest correlation seen in the category of “repetition” (see Table 8). While the results show a similar ranking of participating systems compared

Rating	COV	COV %	COH	COH %	REP	REP %	CON	CON %	REL	REL %
1	42	15.16	4	1.44	23	8.27	6	2.17	4	1.44
2	59	21.30	42	15.11	51	18.35	35	12.64	26	9.35
3	59	21.30	64	23.02	61	21.94	57	20.58	51	18.35
4	71	25.63	81	29.14	71	25.54	82	29.60	83	29.86
5	46	16.61	87	31.29	72	25.90	97	35.02	114	41.01

Table 7: Overview of ratings for Semantic **C**overage, **C**oherence, **R**epetition, **C**onsistency, and **R**elevance. For each evaluation dimension, the first column depicts the absolute number of annotations for a specific score (from 1 to 5) and the second column the percentages.

Dimension	ρ	r
Coverage	.005	-.024
Coherence	.076	.057
Repetition	.117	.025
Consistency	.039	.024
Relevance	.008	.003

Table 8: Correlation between Q + A scores (Hungarian METEOR) and human-rated subset of evidence. We calculate correlation using the Spearman (ρ) and Pearson (r) correlation coefficients.

to human evaluations on the subset, further work is needed to develop scoring methods that align more closely with human assessments of evidence. With that said, overall, the top-ranked teams (based on AVERITEC score) also perform well on human evaluation, while the lower-ranked teams remain similarly positioned, with only minor shifts in their order.⁹ It is important to note that this evaluation was solely based on a small sample of system predictions, and that the results should therefore be taken with a grain of salt.

Human evaluation of evidence predictions offers valuable insights into the limitations of the AVERITEC score, and suggests directions for future research. A notable observation is the discrepancy between human evaluation and the AVERITEC score for some of the highest-ranked samples, such as the examples provided in Table 12 in the appendix. For instance, in row three, the predicted evidence directly contradicts the reference evidence by providing different numbers, yet it receives a high AVERITEC score due to similar wording. Similarly, for the first two rows in Table 12, the semantic coverage score is rated with the second lowest score 1, whereas the average score across all examples is 3, indicating misalignment between the predicted and reference evidence.

⁹See Table 10 in the appendix.

Certain low-ranked examples highlight different challenges (see Table 13). For example, the predicted evidence in the first row received a low AVERITEC score despite receiving the highest score of 5 across all categories in human evaluation. Despite both sets of evidence reaching the same conclusion, the large disparity in answer length and wording leads to a much lower AVERITEC score. The example in the second row, also ranks low according to AVERITEC score, even though it scores high in all categories except for coverage, where it scores 3. Here, both the reference and predicted evidence reach the same verdict, but the predicted evidence supports the claim with different information and wording, resulting in low semantic coverage and a low AVERITEC score.

5 Lessons Learned

Providing a knowledge store rather than requiring participants to rely on a search engine API made the task more accessible. Given the cost of API access, this allowed substantial analysis and work by participants on retrieval. We note that most submissions – 13 of 16 system description papers – used the knowledge store. Nevertheless, because of the size of the knowledge store and the inclusion of distractor documents, the knowledge store did not trivialise the task, and systems relying on search remain competitive and provide unique advantages. Several participants, such as team FZI-WIM, commented on how the two are complementary, and suggested hybrid systems using *both* as a potentially fruitful extension of their systems.

AVERITEC presupposes a strong focus on evidence retrieval. The overall score, as in FEVER (Thorne et al., 2018a), is determined *both* by retrieval performance *and* by veracity prediction performance. In the AVERITEC shared task, participant systems innovated across the pipeline, and all of the top-scoring systems suggest improvements to multiple subtasks of fact-checking.

Team name	0-1000	1000-2215
TUDA_MAI	0.61	0.64
HUMANE	0.55	0.58
CTU AIC	0.45	0.55
Dunamu-ml	0.5	0.5
Papelo	0.49	0.46
UHH	0.41	0.48
SynApSe	0.41	0.43
arioriAveri	0.35	0.42
Data-Wizards	0.32	0.34
MA-Bros-H	0.22	0.31
mitchelldehaven	0.22	0.27
SK_DU	0.2	0.25
UPS	0.15	0.25
FZI-WIM	0.19	0.2
KnowComp	0.19	0.18
IKR3-UNIMIB	0.16	0.2
ngetach	0.12	0.16
VGyasi	0.12	0.12
<i>Baseline</i>	<i>0.11</i>	<i>0.12</i>
InfinityScalers!	0.1	0.07
AYM	0.06	0.06
Factors	0.06	0.04
Average	0.27	0.3

Table 9: AVERITEC scores for different subsections of the dataset. We compute results for the initial test set of 1000 examples collected by Schlichtkrull et al. (2023a), and for the additional 1215 test examples collected for this shared task.

When submitting test set predictions, we required participants to include a field (“*scraped_text*”) for each piece of evidence in their submission, corresponding to the webpage providing backing for that piece of evidence. This enabled us to carry out manual and automatic quality control evaluation verifying that systems do indeed ground their evidence in external sources (see Section 3). This enabled us to detect, for example, if some systems were hallucinating evidence; we did not see any evidence of hallucinated evidence, but we consider guardrails against this crucial. Unfortunately, the inclusion of this field made some submissions substantial in size, as entire webpages were included – up to 2.3gb for the largest submission. Our submission portal, eval.ai, was not able to handle these large files, blocking the portal for all participants during the last few days of the competition. We extended the deadline to compensate.

The scraper we used for the knowledge store (same as in Schlichtkrull et al. (2023a)) to retrieve evidence turned out to be a significant weakness. As some participants noticed, many knowledge store documents are empty. The submission with the best retrieval performance, Dunamu-ml, used a custom scraper, and may have derived significant gains from that choice. We suggest that this may be an interesting area for further research.

During the competition, we identified an issue with the knowledge store data for the last 1215 test examples. Due to an error with date formats, for some claims, web pages published after the claim were included in the knowledge store. This included fact-checking articles, as also mentioned by CTU AIC in their system description paper. As the first 1000 examples were not affected, we computed performance on the first 1000 and last 1215 test examples separately – see Table 9.

As can be seen, the ranking of participants on the two splits is roughly the same – and, indeed, roughly the same as for the entire test set. The second half *was* easier, and many systems perform slightly better there. Somewhat surprisingly, some systems which relied on Google search – specifically, SynApSe – *also* saw a performance gain when measured only on the second split. As such, we do not believe this issue majorly impacted any subset of participants, such as those not relying on the knowledge store. We release an updated knowledge store along with our shared task paper, accessible at <https://fever.ai/dataset/averitec.html>. We have re-compiled the knowledge store with the correct date cutoff, and removed any fact-checking articles that snuck through from the evidence base.

6 Conclusions & Future Work

The AVERITEC shared task attracted submissions from 21 teams, 18 of which outperformed our baseline. The leaderboard was dominated by systems relying on large language models, especially GPT-4o; nevertheless, especially for question generation and retrieval, smaller models – such as LLama-3-8b – also achieved top performance. The winner of the shared task was team TUDA_MAI, which achieved an AVERITEC-score of 63%. In this paper we have analysed the shared task, highlighting aspects of the 16 submitted system description papers, as well as key takeaways from the shared task itself.

The strong performance of the participating teams establishes a firm foundation for automating aspects of real-world fact-checking. The results furthermore indicate clear directions for future work. First, most participating systems – especially for veracity prediction – relied on very large models, such as GPT-4. Further, many of these are blackbox models. These models may be prohibitively expensive for some real-world use cases, e.g., assisting smaller fact-checking organisations (Schlichtkrull et al., 2023b). Given that, we suggest that getting smaller, more efficient models to reach the performance of their larger counterparts may be a fruitful direction for further research. Similarly, we note that performance for most top-scoring systems was much higher on supported and refuted claims, compared to conflicting evidence and not enough evidence. We suggest that leveling this gap is another clear avenue for future improvements.

7 Limitations & Ethics

The datasets and models described in this paper are not intended for truth-telling, e.g. for the design of fully automated content moderation systems. The evidence selection and veracity labels provided in the AVERITEC dataset relate only to the evidence recovered by annotators, and as such are subject to the biases of annotators and journalists. Participant systems, which sought to maximize performance on AVERITEC, may replicate those biases. We furthermore note that shared task leaderboards are a limited representation of real-world task needs, not the least because the test set is static. Acting on veracity estimates arrived at through biased means, including automatically produced ranking decisions for evidence retrieval, risks causing epistemic harm (Schlichtkrull et al., 2023b).

Acknowledgments

Michael, Yulong, Chenxi, Zhenyun, and Andreas received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation programme grant AVeriTeC (Grant agreement No. 865958). Rui is funded by a grant from the Alan Turing Institute and DSO National Laboratories (Singapore). Rami Aly was supported by the Engineering and Physical Sciences Research Council Doctoral Training Partnership (EPSRC). The annotation of the new test set was conducted by a donation from Google.

References

- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. [The fact extraction and VERification over unstructured and structured information \(FEVEROUS\) shared task](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. 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.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Adrien Barbaresi. 2021. [Trafilatura: A web scraping library and command-line tool for text discovery and extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc."
- Svetlana Churina, Anab Maulana Barik, and Saisamarth Rajesh Phaye. 2024. [Improving evidence retrieval on claim verification pipeline through question enrichment](#). In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. [Computational journalism: A call to arms to database researchers](#). In *5th Biennial Conference on Innovative Data Systems Research (CIDR)*.
- John Dougrez-Lewis, Mahmud Elahi Akhter, Yulan He, and Maria Liakata. 2024. [Assessing the reasoning abilities of chatgpt in the context of claim verification](#). *Preprint*, arXiv:2402.10735.
- Andy Dudfield. 2020. [How we’re using AI to scale up global fact checking](#). <https://fullfact.org/>

- <blog/2020/jul/afc-global/>. Accessed: 2023-01-17.
- Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. [Missing counter-evidence renders NLP fact-checking unrealistic for misinformation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- 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.
- H. W. Kuhn. 1955. [The hungarian method for the assignment problem](#). *Naval Research Logistics Quarterly*, 2(1-2):83–97.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*.
- Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracín, Michelle Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn Newman, Gordon Pennycook, Ethan Porter, David G. Rand, David N. Rapp, Jason Reifler, Jon Roozenbeek, Philipp Schmid, Colleen M. Seifert, Gale M. Sinatra, Briony Swire-Thompson, Sander van der Linden, Emily K. Vraga, Thomas J. Wood, and Maria S. Zaragoza. 2020. [Debunking Handbook 2020](#). <https://sks.to/db2020>.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *Preprint*, arXiv:2308.03281.
- Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang, and Yangqiu Song. 2024a. GProofT: A multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Jin Liu, Steffen Thoma, and Achim Rettinger. 2024b. FZI-WIM at averitec shared task: Real-world fact-checking with question answering. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Christopher Malon. 2021. [Team papelo at FEVEROUS: Multi-hop evidence pursuit](#). In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49, Dominican Republic. Association for Computational Linguistics.
- Christopher Malon. 2024. Multi-hop evidence pursuit meets the web: Team papelo at FEVER 2024. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Shrikant Malviya and Stamos Katsigiannis. 2024. SK_DU team: Cross-encoder based evidence retrieval and question generation with improved prompt for the AVeriTeC shared task. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. [Generation-augmented retrieval for open-domain question answering](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xixi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Sebastião Miranda, Andreas Vlachos, David Nogueira, Andrew Secker, Afonso Mendes, Rebecca Garrett, Jeffrey J Mitchell, and Zita Marinho. 2019. [Automated fact checking in the news room](#). In *The Web Conference 2019*, pages 3579–3583, United States. Association for Computing Machinery (ACM). 2019 World Wide Web Conference, WWW 2019 ; Conference date: 13-05-2019 Through 17-05-2019.
- Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. [Looking beyond sentence-level natural language inference for question answering and text summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

- Mohammad Ghiasvand Mohammadkhani, Ali Ghiasvand Mohammadkhani, and Hamid Beigy. 2024. Zero-shot learning and key points are all you need for automated fact-checking. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Yuki Momii, Tetsuya Takiguchi, and Yasuo Ariki. 2024. RAG-fusion based information retrieval for fact-checking. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. **MTEB: Massive text embedding benchmark**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. **Automated fact-checking for assisting human fact-checkers**. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Adjali Omar. 2024. Exploring retrieval augmented generation for real-world claim verification. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. **Varifocal question generation for fact-checking**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Heesoo Park, Dongjun Lee, Jaehyuk Kim, Choongwon Park, and Changhwa Park. 2024. Dunamu-ml’s submissions on AVeriTeC shared task. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Karl Pearson. 1896. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318.
- Stephen Robertson and Hugo Zaragoza. 2009. **The probabilistic relevance framework: Bm25 and beyond**. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Mark Rothmel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. InFact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023a. **Averitec: A dataset for real-world claim verification with evidence from the web**. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.
- Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023b. **The intended uses of automated fact-checking artefacts: Why, how and who**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. **Get your vitamin C! robust fact verification with contrastive evidence**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.
- Özge Sevgili, Irina Nikishina, Seid Muhie Yimam, Martin Semmann, and Chris Biemann. 2024. UHH at AVeriTeC: RAG for fact-checking with real-world claims. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Ronit Singhal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- C. Spearman. 1987. **The proof and measurement of association between two things**. *The American Journal of Psychology*, 100(3/4):441–471.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. **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.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. **The fact extraction and VERification (FEVER) shared task**. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.
- Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Nicolò Urbani, Sandip Modha, and Gabriella Pasi. 2024. Retrieving semantics for fact-checking: A comparative approach using CQ (claim to question) & aq (answer to question). In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. **Fact checking: Task definition and dataset construction**. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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

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

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. The herd of open llms for verifying real-world claims. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. **mgte: Generalized long-context text representation and reranking models for multilingual text retrieval**. *Preprint*, arXiv:2407.19669.

A Search Queries for Knowledge Store Generation

When creating the knowledge stores for the train, development, and test set, we used a series of search query generation strategies. An overview can be seen in Table 11. We note that some of these rely on information not available normally to participants, such as the gold question-answer pairs. We note that, despite this, systems not relying on the knowledge store, such as Papelo, were competitive.

B Human Evaluation

We carried out human evaluation of the submitted test set predictions. Below in Figures 2-9, we include screenshots of the interface used by annotators. We also include, in Tables 12 and 13, instructive examples from the human evaluation.


Source	Score Coverage
CTU AIC	4.1
TUDA_MAI	4.1
SynApSe	3.8
Dunamu-ML	3.5
MA-Bros-H	3.4
Factors	3.3
Data-Wizards	3.2
UHH	3.2
mitchelldehaven	3.1
SK_DU	3.1
IKR3-UNIMIB	3.1
FZI-WIM	2.9
InfinityScalers!	2.9
arioriAveri	2.9
HUMANE	2.8
Papelo	2.8
KnowComp	2.8
UPS	2.4
VGyasi	2.3
AYM	2.3
ngetach	2.0

Table 10: Average scores assigned to evidence samples from different participating teams for the semantic coverage category, based on human evaluation.

Query type	Description
Generated questions	<i>Questions are generated with gpt-3.5-turbo based on the claim. Three claim-question pairs from the training set are used as in-context examples.</i>
Generated background queries	<i>Queries are generated with gpt-3.5-turbo based on the claim. The prompt focuses on background information, such as details about entities in the claim. Three manually constructed claim-query pairs are used as in-context examples.</i>
Generated provenance queries	<i>Queries are generated with gpt-3.5-turbo based on the claim. The prompt focuses on information necessary to establish provenance, such as whether the claim source is a satire site. Three manually constructed claim-query pairs are used as in-context examples.</i>
Claim named entities	<i>Named entities from the claim are extracted and used as search queries. One query for each entity is constructed, along with one query containing all entities.</i>
Most similar gold evidence	<i>The most similar paragraph in the gold evidence document is selected using BM25, and used as a search query.</i>
Gold URL generated questions	<i>Queries are generated with gpt-3.5-turbo based on the URL of the gold evidence. The prompt tried to generate questions that would retrieve the URL in question. Three manually constructed URL-query pairs are used as in-context examples.</i>
Different event same entity	<i>Queries are generated with gpt-3.5-turbo based on the named entities in the claim. The prompt focuses on different events involving some of the same entities. Results are used as distractors to make the retrieval task harder.</i>
Similar entities	<i>Queries are generated with gpt-3.5-turbo based on the claim. The prompt replaces entities in the claim with other similar entities, such as changing one city to another. Results are used as distractors to make the retrieval task harder.</i>
Gold questions	<i>Gold questions used verbatim as search queries.</i>
Claim + gold question	<i>Gold questions used verbatim as search queries. The claim is prepended, processed as in Schlichtkrull et al. (2023a).</i>
Rephrased gold questions	<i>Gold questions are rephrased using gpt-3.5-turbo, and then input as search queries.</i>
Gold answers	<i>Gold questions used verbatim as search queries.</i>
Rephrased gold answers	<i>Gold answers are rephrased using gpt-3.5-turbo, and then input as search queries.</i>

Table 11: Queries input to the Google Search API for each claim in order to build the knowledge store. Following [Schlichtkrull et al. \(2023a\)](#), we restrict search results to documents published before the claim. For each claim, we also extend the knowledge store with the corresponding gold evidence documents.

Evidence Evaluation for AVERITEC System Predictions

mubashara.ak@gmail.com [Switch account](#) 

Intro

Thank you for helping to evaluate the AVeriTeC shared task submissions!

For the shared task (<https://fever.ai/task.html>), many teams have submitted predictions, including claim labels and evidence. Your task is to rate these submissions to support a detailed study of the results.

Please find the selected submissions you need to rate in this folder (select the file named with your team name):

Each example provided for evaluation consists of the following fields:

1. The **claim ID**
2. The **claim**
3. The **predicted label**
4. The **predicted evidence** extracted from a shared task submission (incl., the scraped text if available)
5. The **reference evidence** for the same claim (i.e., the "gold" evidence)

[Back](#) [Next](#) [Clear form](#)

Figure 2: Platform for human evaluation of retrieved evidence from participating systems.

Claim Verdict based on Predicted Evidence

On this page, please do the following:

1. Check if the **predicted evidence** contains major errors that warrant skipping the example.
2. Label the claim based on the **predicted evidence** as one of the following:
 - o **Supported**
 - o **Refuted**
 - o **Not Enough Evidence**
 - o **Conflicting Evidence/Cherry-picking**

Enter [Claim ID] below: *

Your answer _____

Enter [Claim] below: *

Your answer _____

Enter the [Predicted Evidence] text below: *

Your answer _____

1. Does the **predicted evidence** contain any of the following three major errors? If *
yes, which of the following holds for the **predicted evidence**?

- Yes, the evidence is ENTIRELY EMPTY
- Yes, the evidence is NOT UNDERSTANDABLE AT ALL
- Yes, the evidence is COMPLETELY IRRELEVANT to the claim
- No major errors. AT LEAST SOME PART of the evidence is non-empty, understandable, and related to the claim.

Figure 3: Platform for human evaluation of retrieved evidence from participating systems.

For the following question:
If you selected "Yes, ..." for the last question (first three options), please skip the question below and submit your response.

If you selected the last option, "No major errors. [...]", proceed to the next question. For the next question, review 1.) the claim and 2.) the **predicted evidence**.

2. Now, decide if the **claim** is (a.) **supported** by the **predicted evidence**, (b.) **refuted**, (c.) **not enough evidence** is given (if there isn't sufficient evidence to either support or refute it), (d.) **conflicting evidence/cherry-picking** (if the claim has both supporting and refuting evidence).

a. supported

b. refuted

c. not enough information

d. conflicting/cherry-picking

3. If you selected options a.) supported, b.) refuted, or d.) conflicting/cherry-picking, please copy from the field "**scraped text**" (if it is available) the text which supports your decision.

Your answer

[Back](#) [Next](#) [Clear form](#)

Figure 4: Platform for human evaluation of retrieved evidence from participating systems.

Rating of Predicted Evidence

Rate the predicted evidence by answering the questions below.

For the first question, you will need to compare the **predicted evidence** to the **reference evidence**.

1. Semantic Coverage

Evaluate **how much of the reference evidence is covered by the predicted evidence**. Compare the two based on their content (e.g., meaning, the extent to which entities in the reference evidence are represented in the predicted evidence, etc.).

1 score: The predicted evidence covers none of the reference evidence.

2 scores: Very little of the reference evidence is covered.

3 scores: Approximately half of the reference evidence is covered.

4 scores: Most of the reference evidence is covered.

5 scores: Everything mentioned in the reference evidence is covered by the predicted evidence.

1 2 3 4 5

Figure 5: Platform for human evaluation of retrieved evidence from participating systems.

For the questions below, you will only need to look at the **predicted evidence!**

2. Coherence

Evaluate the coherence of the **predicted evidence** by assessing if all sentences are logically and meaningfully connected to one another, and if the evidence makes sense as a whole.

1 score: Not coherent at all.

2 scores: Most of the text is incoherent, with sentences disconnected and the overall meaning unclear.

3 scores: Approximately half of the evidence is coherent, while the rest is not.

4 scores: Almost every sentence is coherent, and the evidence mostly makes sense as a whole, with some minor mistakes.

5 scores: Very coherent; the entire text forms a unified and logical body.

1

2

3

4

5

Figure 6: Platform for human evaluation of retrieved evidence from participating systems.

3. Repetition

Evaluate the **predicted evidence** for any repetition.

1 score: A lot of repetition; most of the evidence text is redundant.

2 scores: A significant portion of the text repeats the same information.

3 scores: Approximately half of the text is repeated content.

4 scores: Minor repetitions in the text.

5 scores: No repetition at all.

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 7: Platform for human evaluation of retrieved evidence from participating systems.

4. Consistency

Evaluate the consistency of the **predicted evidence** in the information it provides.

1 score: Not consistent at all; contains a lot of conflicting and/or illogical information.

2 scores: Most of the evidence is inconsistent, with major parts that conflict or are illogical.

3 scores: Approximately half of the evidence is consistent, but there are significant conflicts or illogical information.

4 scores: The evidence is mostly consistent, with a few minor issues such as confusion of dates, names, or other details.

5 scores: The evidence is very consistent, with no conflicting or illogical information.

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 8: Platform for human evaluation of retrieved evidence from participating systems.

5. Relevance to Claim

Evaluate how relevant the **predicted evidence** is to the claim.

1 score: Not relevant at all; the evidence does not relate to the claim in any meaningful way.

2 scores: Mostly irrelevant, with only a small portion of the evidence having minor relevance to the claim.

3 scores: Approximately half of the evidence is relevant to verifying the claim, while the rest is redundant or unrelated.

4 scores: Most of the evidence is relevant, with some minor irrelevant or redundant parts.

5 scores: Very relevant; the evidence is entirely focused on verifying the claim without any irrelevant information.

1 2 3 4 5

[Back](#) [Submit](#) [Clear form](#)

Figure 9: Platform for human evaluation of retrieved evidence from participating systems.

	claim	predicted evidence	label human eval	semantic coverage	coherence	redundancy	consistency	relevance	predicted label	reference evidence
1	New changes to the Highway Code in Great Britain states drivers should give five feet when overtaking cyclists.	Question: Should drivers give five feet when overtaking bicyclists? Answer: No answer could be found. Question: What is the Highway Code in Great Britain? Answer: No answer could be found.	not enough information	1	5	1	5	5	Not Enough Evidence	Question: What is 5 feet in metres? Answer: 1.5 metres. Question: What is the new overtaking distance between drivers and cyclists in the Highway Code? Answer: leave at least 1.5 metres when overtaking cyclists. Question: At what maximum speed is the new 1.5-distance rule applicable? Answer: at speeds of up to 30mph. Question: Should drivers leave more space if driving faster? Answer: Yes. On a PDF for the 'Table of Changes to the Highway Code' it states, "Give them more space when overtaking at higher speeds."
2	Tweet portrays Ravish Kumar targeting AAP ahead of Punjab elections	Question: When were the Punjab elections? Answer: No answer could be found. Question: When was the tweet made? Answer: The tweet was allegedly made by Ravish Kumar in the midst of the Punjab election. Question: Is Ravish Kumar an AAP supporter? Answer: No answer could be found.	not enough information	1	2	5	3	2	Supported	Question: What does Ravish Kumar think of the AAP? Answer: No answer could be found. Question: Did Ravish Kumar tweet about the AAP? Answer: No answer could be found. Question: Has Ravish Kumar been misquoted on political matters? Answer: Senior Congress leader Digvijaya Singh apologised to NDTV anchor Ravish Kumar today after sharing a video on Twitter that incorrectly suggested the latter had abused Prime Minister Narendra Modi in a public speech on Wednesday.
3	In 2010, a newly qualified teacher's salary in the UK was £21,500 and in 2018 it was £22,000.	Question: What was the starting salary for a newly qualified teacher in 2010? Answer: £21,500. Question: What was the starting salary for a newly qualified teacher in 2018? Answer: £22,000. Question: What was the starting salary for a newly qualified teacher in 2010 in 2018 prices? Answer: £25,400. Question: What was the starting salary for a newly qualified teacher in 2018 in 2010 prices? Answer: £23,000.	supported	5	4	1	2	4	Supported	Question: What was the salary of a newly qualified teacher in the UK in 2010? Answer: £21,588. Question: What was the salary of a newly qualified teacher in the UK in 2018? Answer: £23,720. Question: What was the salary of new teachers outside of London in 2010? Answer: £21,588. Question: What was salary of new teachers outside of London in 2018? Answer: Minimum starting salaries for classroom teachers will increase by between £653 (rest of England) and £816 (inner London). This means the minimum starting salary for a qualified teacher in 2019/20 will rise from £23,720 to £24,373 outside of London and from £29,664 to £30,480 in inner London.

Table 12: Examples for system predictions which are highly-ranked by the Averitec score.

	claim	predicted evidence	label human eval	semantic coverage	coherence	redundancy	consistency	relevance	predicted label	reference evidence
1	The reporting environment for foreign correspondents in China is open and free. Communication channels between foreign journalists in China and competent Chinese authorities are open and unfettered.	Question: What is the working environment for foreign journalists in China? Answer: Foreign correspondents in China have reported higher levels of concern about surveillance and invasion of privacy, as well as greater pressure by overseas Chinese officials on media organization headquarters. Nearly half of surveyed correspondents said they experienced interference, harassment, and physical violence. [...]	refuted	5	5	5	5	5	Refuted	Question: Does the Chinese government place restrictions on journalist content? Answer: Yes. The Propaganda Department of the Chinese Communist Party sends a detailed notice to all media every day that includes editorial guidelines and censored topics.
2	Philippines Senator Panfilo Lacson changes his opinion and says he sees the sense in the Department of National Defense's (DND) unilateral termination of its 1989 accord with the University of the Philippines (UP).	Question: What was Philippines Senator Panfilo Lacson's previous stance on the Department of National Defense's (DND) termination of its 1989 accord with the University of the Philippines (UP)? Answer: Philippines Senator Panfilo Lacson's previous stance on the Department of National Defense's (DND) termination of its 1989 accord with the University of the Philippines (UP) was critical. [...]	supported	3	5	5	5	4	Supported	Question: What was his initial view on the DND and University of the Philippines longstanding arrangement? Answer: There's sense in the move taken by the President and Sec. Lorenzana because during our red-tagging hearings in the Senate, it was established that—not only UP in particular, other universities—talagang 'yung recruitment nanggagaling dun sa mga estudyante up to the point that they're being killed in encounters, he told ANC's Headstart.

Table 13: Examples for system predictions which are low-ranked by the Averitec score.