

The 2nd Automated Verification of Textual Claims (AVeriTeC) Shared Task: Open-weights, Reproducible and Efficient Systems

Mubashara Akhtar¹, Rami Aly², Yulong Chen², Zhenyun Deng²,
Michael Schlichtkrull^{2,3}, Chenxi Whitehouse^{2,4}, Andreas Vlachos²

¹ETH Zurich, ²University of Cambridge, ³Queen Mary University of London, ⁴Meta,
mubashara.akhtar@ai.ethz.ch, {rmya2,yc632,zd302,cj507,av308}@cam.ac.uk
m.schlichtkrull@qmul.ac.uk

Abstract

In the First Automated Verification of Textual Claims (AVERITEC) shared task, participating teams developed systems that for each claim retrieve evidence from the web and predict its veracity. While there was progress in automated fact-checking for real-world claims, the majority of the systems proposed relied on large closed-weights language models, which rendered them expensive to run and less reproducible. To ameliorate this issue, in this year’s edition of the AVERITEC shared task, we required system to use only open-weights models that could be run using a single GPU with 23GBs of RAM, and that systems should take one minute or less to return verdicts accompanied by evidence retrieved from a pre-compiled knowledge store. The shared task received 7 submissions; 6 of which exceeded the accuracy of our baseline on the test set, while they ran in under a minute per claim on the hardware we had specified. The winning team was CTU AIC with an AVERITEC score of 33.17%. In this paper we describe the shared task in detail and highlight key findings.

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). While there had been progress on many benchmarks, these were limited in their ability to measure progress in terms of evidence retrieval. For example, FEVER (Thorne et al., 2018) relied on Wikipedia as its only source of evidence, in addition to consisting of purpose-made rather than real-world claims. Liar Liar Pants on Fire (Wang, 2017) consists of real-world claims but it has no

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.

evidence annotated to evaluate retrieval, while in MultiFC (Augenstein et al., 2019) the evidence is annotated automatically and thus cannot be relied upon for evaluation (Glockner et al., 2022).

The recently proposed AVERITEC dataset (Schlichtkrull et al., 2023a) addressed these limitations. It consists of real-world claims where the evidence is manually annotated in the form of questions and answers sourced from the Web (see Figure 1 for an example). This evidence had to be available before the claim was made, and was additionally verified to adequately support the verdict, thus avoiding the issues of temporal leakage and evidence insufficiency identified in earlier datasets (Ousidhoum et al., 2022; Glockner et al., 2022).

In the first AVERITEC shared task (Schlichtkrull et al., 2024), a number of systems were proposed that substantially improved the results on the task compared to the baseline proposed with the dataset. However, most relied on closed-weights large language models (LLMs) with substantial (and unverified) parameter counts, including the top-performing system (Rothermel et al., 2024). While the tremendous progress achieved was inspiring, and the evaluation of commercial LLMs in the context of automated fact-checking was of high practical significance, this result has a number of shortcomings from a research perspective.

First, the systems rely on external LLMs via APIs that have no control over them; thus it is difficult to reproduce their results, and as LLM provider change their offering overtime, it becomes impossible. Second, the costs of developing such approaches could be substantial depending on the size of the LLMs used. Thus, the financial ability to use such commercial services affected the chances of achieving good results on the shared task. More problematically, journalistic fact-checkers, the most common intended user for automated fact-checking tools (Schlichtkrull et al., 2023b), often operate under strong resource constraints and cite cost as a major deciding factor for the adoption of technologies (Warren et al., 2025). Expensive models may as such not be able to match their desired real-world function. Last but not least, systems relying on external LLMs can be less practical to use in real-world contexts, where latency due to network limitations and/or privacy concerns are important considerations.

For these reasons, in this second edition of the AVERITEC shared task, we decided to focus on open-weights, reproducible and efficient systems. Participating systems were constrained to run using a single GPU taking a maximum of one minute per claim to return their verdicts. This runtime included retrieving evidence from a pre-compiled knowledge store consisting of documents returned by a commercial search engine. Note that participants were allowed to use larger and/or close-weights LLMs for training, these restrictions only applied to inference. Similar to the first edition of the shared task, this knowledge store contains the manually annotated evidence that systems need to return with their verdicts, but also a lot of other related search results. It was preferred over offering systems the option to access a commercial search engine directly as it is free to use by the participants

(and it was the preferred option by most of them in the first edition), but also that participating systems did not need to access any resources beyond the knowledge store and models running locally. To ensure that all participating teams adhered to these restrictions, they were asked to submit their systems to the organisers to run on the test data in order to produce the final results.

We also improved on the automatic evaluation of evidence retrieval, which was found to have very low correlation with human evaluation in the first edition of the shared task. Instead of relying on the originally proposed token-matching evaluation of Schlichtkrull et al. (2023a), we adopted the recently proposed Ev²R prompt-based LLM approach of Akhtar et al. (2024) which was shown to have stronger correlation with human evaluation. This allowed for more reliable evaluation of participating systems, since accuracy points are awarded conditionally on retrieving appropriate evidence. Finally, we released a new test set with more recent claims, thus reducing the possibility that they were used in the training of LLMs.

We find that all seven participating teams delivered systems that adhered to the requirements for open-weights, reproducible and efficient systems, making use of language models up to 14B parameters. Fine-tuning was rarely used, relying mostly on few-shot in-context learning. In retrieval, they often proposed hybrid approaches combining dense embeddings with BM25. Overall, the best system was submitted by team CTU AIC Ullrich and Drchal (2025) that achieved 0.3317 AVERITEC score, which awards accuracy points only when the evidence retrieved is considered adequate.

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, but most participants found it beneficial to follow it. 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*.

2.1 Dataset

Similarly to the shared task of previous years, we ask participants to train and validate their system on the public AVERITEC dataset and evaluate their performance on our new test set (2025). The 2025 test set consists of 1,000 instances, which temporally succeed the previous data (original AVERITEC and 2024 test set).

Annotation of 2025 New Test Set Following AVERITEC (Schlichtkrull et al., 2023a), we first collect fact-checking articles from ClaimReview and conduct a five-phase annotation. Please note that each instance is annotated by different annotators at each phase.

In particular, in phase 1 (P1), annotators are asked to identify the main claims from a fact-checking article, extract corresponding meta-data such as the claim speaker and claim date, and decontextualise the claim to make it context-independent. In P2, given a decontextualised claim, annotators propose questions that help to fact-check the claim, answer the question by finding relevant information from the fact-checking article and online source, and finally make a verdict for the claim based on the QA pairs. In P3, presented with only the QA pairs and the decontextualised claim, annotators assign a verdict label and make a justification for their choice. For each instance, if the verdict labels given by P2 and P3 annotators are identical, we regard the QA annotation and the claim as disambiguated, informative and sufficient for the verdict predication, and include it in our resulting test set. Otherwise, this instance proceeds to the P4 and P5 annotation, which consist of a second round of P2 and P3 annotation, respectively. Similarly, we examine the verdicts of each claim given by P4 and P5 annotators: if the verdict labels are identical, we include the instance in our resulting test set; otherwise, we discard it. In this way, we collect 1,000 instances for our 2025 test set, where each is annotated with a normalised claim, meta-data, QA pairs, a verdict label, and a justification.

We conduct a training and evaluation procedure to select qualified annotators. Before the formal annotation, all annotators are required to complete training on 10 instances for each of the P1, P2, and P3 tasks, respectively. Those training instances are randomly selected from the 2024 test set. All annotators are required to meet the basic performance criteria: (1) over 70% F -1 score for both claim type classification and fact-checking strategy clas-

sification; (2) an average of more than 2 QA pairs per claim; (3) over 50% accuracy of verdict prediction. Finally, 8 out of 9 annotators are selected for the formal annotation for the 2025 test set.

Comparison between the Previous Datasets and 2025 Test Set We present the data statistics of the 2025 test set in Table 1. For comparison, we also show the results of the 2023 AVERITEC dataset and the 2024 test set.

The 2025 test set (from Jan 2024 to Dec 2024) is more temporally removed from the training set compared with the 2023 dataset and the 2024 test set, indicating a greater domain shift. The average number of questions per claim in the 2025 test set is comparable to the 2024 test set while being higher than in the 2023 dataset (e.g., 2025: 2.79; 2024: 2.89; 2023: 2.60/2.57/2.57). Moreover, the 2025 test set includes more numerical claims (38.8%), which are more straightforward to verify, but fewer causal claims (8.1%), which are typically more challenging. These distributions also reflect on the fact-checking strategies, where there are more numerical comparisons (30.8%) in the 2025 test set.

In addition to the above observations, we find that the distributions across different sets show similar trends. In terms of label distribution, the Refuted label consistently accounts for the largest proportion, while Conflicting and Not Enough Evidence remain greatly fewer. Regarding claim type distribution, the Event/Property Claim is the most common, while the Position Statement is the least. For fact-checking strategies, the Written Evidence consistently dominates across all sets.

2.2 Knowledge Store

To ensure fair comparison, support reproducibility, and reduce engineering and computational costs, we provide a corresponding knowledge store for the 2025 test set. For each claim, the knowledge store includes a set of potentially relevant documents for fact-checking each claim.

For each data store, we include gold documents, which are used for our annotation, and additional documents retrieved by Google search. In particular, to generate queries for Google search, we use ChatGPT¹ to generate a set of queries based on the claim, gold annotated questions, and gold annotated answers. We also include a variety of distractor queries by changing the named entities,

¹We use gpt-3.5-turbo.

Split	Train (2023)	Dev (2023)	Test (2023)
Claims	3,068	500	1,000
Question / Claim	2.60	2.57	2.57
End date	25-08-2020	31-10-2020	22-12-2021
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
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
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

Split	Test (2024)	Test (2025)
Claims	1,215	1,000
Question / Claim	2.89	2.79
End date	13-08-2023	19-12-2024
Labels (S/R/C/N)	17.3/66.5/4.1/12.1	22.2/71.9/1.7/4.2
Types (PS/NC/EPC/QV/CC)	3.5/24.3/71.9/5.2/16.1	2.6/38.8/68/9/4.3/8.1
Strategies (WE/NCP/FR/EC/SS)	82.4/22.6/10.0/37.6/4.0	88.8/30.8/7.8/30.0/5.6

Table 1: Statistics for the 2023 dataset, and 2024 and 2025 test sets. 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). For simplicity, we exclude strategies with very low frequencies, such as Geo-location (0.3%). Please note that a single claim can correspond to multiple claim types and fact-checking strategies; therefore, the proportions do not necessarily sum to 100%.

dates, and events in the claim. We present our detailed query information in Appendix A. We collect the URLs returned by the first page of the Google search, and only include those URLs which are temporarily available before the claim is made. Finally, the deduplicated and shuffled URLs result in the data store for each claim. We further scrape the text from each URL using *trafilatura* (Barbresi, 2021).

For the 2025 test data store, we have 1,018,800 URLs and 2,506,398,451 tokens in total. In particular, for each claim, there are 1019 URLs on average, where 593 are associated with valid scraped texts. The average tokens are 2,506,398 for each claim and 4,227 for each document, respectively. The most common domains include National Library of Medicine, Reddit, ScienceDirect, Wikipedia, BBC, the New York Times and CNN.

2.3 Baseline

The baseline closely follows the HerO system (Yoon et al., 2024). HerO achieved the second place in the AVeriTeC shared task (Schlichtkrull et al., 2024), demonstrating that open LLMs can effectively verify real-world claims without relying on proprietary models. HerO uses publicly available LLMs in a three-step verification pipeline: (i) evidence retrieval by combining hypothetical document generation via an LLM, BM25 retrieval, and a cross-attention re-ranker (Meng et al., 2024); (ii) question generation where an LLM creates veri-

fying questions conditioned on each piece of the evidence; and (iii) veracity prediction by using a fine-tuned LLM to jointly generate explanations and the final verdict labels.

Our baseline modifies the original HerO implementation with a focus on computational efficiency to ensure that the system runs within this shared task’s time constraints. Instead of using Llama-3.1-70B (Grattafiori et al., 2024) across components, the baseline uses the Llama-3.1-8B variant (a fine-tuned Llama-3.1-8B veracity prediction model was also provided by Yoon et al. (2024)). Since evidence retrieval is the most expensive step of HerO’s inference pipeline, the baseline additionally incorporates retrieval cutoffs and heuristics, limiting the number of sentences for BM25 retrieval to 5000, and for reranking to 500. Finally, the runtime was further improved by adding typical efficiency optimizations, such as batch processing and multi-threading.

2.4 Measuring Reproducibility & Efficiency

To ensure the reproducibility of shared task systems, all systems were executed on a standardized virtual machine during inference on the test set by the organizers. To this end, all shared task teams were required to provide reproducible code with clear installation and execution instructions.

A system is considered reproducible if it runs during inference on the VM without making any external API calls, whether to large language mod-

els (LLMs) or to retrieval engines such as Google Search. Consequently, closed-weight LLMs cannot be used during inference. In contrast, open-weight and open-data language models are allowed as long as they run locally on the VM. Note that participants were allowed to use larger or closed-weight LLMs during training.

The virtual machine was an AWS g5.2xlarge EC2 instance with an Nvidia A10G GPU with 23GB memory, 8 vCPUs, 32GB RAM, and 450GB of storage. To ensure compatibility with the VM, participants could test their systems using either a provided Docker image that matched the evaluation environment or by configuring an identical AWS instance via the specified AMI.

The efficiency of shared task systems was measured by setting an upper limit to the inference runtime on the 1000 claims of the test set. A system was expected to process the entire test set on the virtual machine in 16 hours and 40 minutes, averaging 1 minute per claim. This runtime limit does not include the downloading of data, models, or retrieval indices. Outputs produced by a system beyond that time constraint are not considered. Moreover, systems were allowed to process all claims for a given component of the verification pipeline before proceeding to the next component. This approach reduces the impact of loading and unloading models from memory that would occur if each claim were processed individually.²

Reproducibility and efficiency are a binary pass/fail requirement for successful shared task submissions. We do not use them as a metric for ranking successful shared task systems.

2.5 Evaluation

Following established practise in previous work (Thorne et al., 2018; Schlichtkrull et al., 2023a), including the first AVeriTeC shared task (Schlichtkrull et al., 2024), we evaluate verdict accuracy conditional on sufficient evidence having been retrieved. We report three metrics: **Q score**, representing question quality regardless of found evidence; **Q+A score**, representing the quality of evidence as questions *and* answers; and **AVERITEC**, on which systems are scored with verdict accuracy for claims where Q+A score is above a certain threshold t , and 0 otherwise.

The evidence in this year’s AVeriTeC shared

²This relaxation creates an admittedly artificial setting, as it would require all users to wait for all claims to be processed before receiving a response.

task is retrieved from a knowledge store compiled from a range of internet sources (see Sec. 2.2). The AVeriTeC metrics used in the previous year’s shared task (Schlichtkrull et al., 2024) relied on approximate matching using the annotated evidence and the token-matching metrics METEOR (Banerjee and Lavie, 2005). However, this approach was highly sensitive to surface forms and resulted in penalising alternative, but valid evidence paths. For example, both “*Where did South Africa rank in alcohol consumption? In 2016, South Africa ranked ninth out of 53 African countries.*” and “*What’s the average alcohol consumption per person in South Africa? 7.1 litres.*” may both be valid ways of establishing the relative levels of alcohol consumption between South Africa and other countries. However, the token-level overlap between both evidence is low and may result in a higher METEOR score for one evidence alternative compared to the other.

Thus we decided to use Ev²R (Akhtar et al., 2024) for evaluation. Ev²R (Akhtar et al., 2024) is a prompt-based LLM-as-judge approach that assesses the quality of retrieved evidence by decomposing both the retrieved and reference evidence into atomic facts before comparing them to evaluate factual consistency and coverage. It outperforms traditional metrics in alignment with human judgments and robustness to adversarial perturbations. Ev²R is inspired by FactScore (Min et al., 2023), but adapts its approach to better reflect evidence evaluation, providing both a precision and a recall score. Precision measures the accuracy of the retrieved evidence, while recall assesses the completeness of the retrieved evidence in relation to the gold standard. The scorer first splits the retrieved evidence \hat{E} and reference evidence E into atomic facts, $A_{\hat{E}}$ and A_E respectively. To calculate the precision score it evaluates whether each individual fact $a_{\hat{E}} \in A_{\hat{E}}$ of the retrieved evidence is supported by the reference evidence E . The precision score s_{prec} is defined as the ratio of facts supported by the reference evidence:

$$s_{prec} = \frac{1}{|A_{\hat{E}}|} \sum_{a_{\hat{E}} \in A_{\hat{E}}} I[a_{\hat{E}} \text{ supported by } E]$$

The scorer iterates over each fact ($a_{\hat{E}} \in A_{\hat{E}}$) for which the indicator function ($I[a_{\hat{E}} \text{ supported by } E]$) returns 1 if the fact $a_{\hat{E}}$ is supported by the reference evidence E and 0 otherwise. For calculating the recall score, the scorer evaluates whether each atomic fact of the

reference evidence ($a_E \in A_E$) is supported by the retrieved evidence, i.e., measuring the extend to which the retrieved evidence covers the content of the reference evidence:

$$s_{recall} = \frac{1}{|A_E|} \sum_{a_E \in A_E} \mathbb{I}[a_E \text{ supported by } \hat{E}]$$

Akhtar et al. (2024) assess the validity of the scorer by evaluating its alignment with human ratings and testing its robustness through a set of perturbation experiments that systematically assess the scorer on various dimensions, such as its sensitiveness to variant changes in the evidence text, fluency, noise, etc.

Following the first AVERITEC shared task (Schlichtkrull et al., 2024), we evaluate evidence using only the recall component of the metric. By doing so we avoid penalising systems for adding additional evidence which annotators did not find necessary, such as background context. We only consider the first 10 questions generated by each system, so as to avoid rewarding sheer volume. We then calculate total AVERITEC score as verdict accuracy given that $s_{recall} > t$, where we choose $t = 0.5$ so as to ensure high agreement on the 100 double-annotated AVERITEC claims following the methodology discussed in Schlichtkrull et al. (2023a).

3 Results

The results for the shared task are shown in Table 2. We received seven fully reproducible systems. This section discusses our findings on reproducibility, efficiency, and general observations on the techniques used by the participating teams. We provide a high-level overview of the model components used by systems in Table 3. For detailed descriptions of any particular system, we refer to each team’s system description paper. In line with the theme of this shared task, every team has made their codebase publicly available.

Reproducibility We received a total of eight system submissions. One system failed to run on the VM due to syntax errors, missing installation instructions, and hardcoded file paths. Of the seven reproducible systems, two were submitted as Docker images and five as ZIP files. All systems needed manual intervention to run on the virtual server. Common issues were Docker permission errors, dependency installation failures (e.g.,

llama.cpp), GPU memory crashes, and misconfigured shell scripts. Only memory crashes occurred during runtime; all other errors were resolved within 4 hours before system execution. Overall, the encountered issues are expected for early-stage open-source codebases.

We added several diagnostic measures to assess a system’s reproducibility. First, we monitored traffic and non-local API calls. Second, we tested each system on a subset of 99 claims not included in the test set (but included in the knowledge store, in case of pre-computed indices) to verify that systems were not hardcoded to specific test examples and could handle arbitrary claims.

Efficiency The average runtime per claim for each system is shown in Table 2. All systems successfully stayed below the established limit of 1 minute per claim on average. Teams achieved this through model selection and efficiency implementation improvements. The components used by each team, along with inference engines and efficiency-focused designs, are summarized in Table 3. Five out of seven systems use for LLM inference vLLM (Kwon et al., 2023), following the baseline. Team EFC uses llama.cpp³ and Team CTU AIC uses Ollama⁴, a wrapper around llama.cpp.

To improve retrieval efficiency beyond the baseline’s improvements, systems CTU AIC, HUMANE, FZIGOT, and OldJoe used pre-computed indices of dense vector representations. Teams Yellow Flash, EFC, and Checkmate chunked evidence sentences into larger segments before applying a sparse BM25 retriever, reducing the number of chunks considered by the BM25 module in Team EFC’s case from 5000 to 1500.

Due to VM resource constraints, most teams used smaller models for both retrieval and veracity prediction than in the first AVeriTeC Shared Task (Schlichtkrull et al., 2024). For instance, Team HUMANE used an 8B model for their retrieval pipeline instead of the 70B model from the first shared task to fit within the 23GB RAM of the A10G GPU. Subsequently, most teams used quantization to either fit larger models onto the GPU and to reduce inference runtime. Teams HUMANE, FZIGOT, and OldJoe used Activation-aware Weight Quantization (Lin et al., 2024), Teams Yellow Flash and Checkmate used OPTQ (Frantar et al., 2023), and Team EFC used GGUF

³<https://github.com/ggml-org/llama.cpp>

⁴<https://github.com/ollama/ollama>

#	Team Name	Time per Claim (s)	Ev2R Recall		AVERITeC Score
			Q only	Q + A	
1	CTU AIC (Ullrich and Drchal, 2025)	53.67	0.2003 _{0.007}	0.4774 _{0.004}	0.3317 _{0.002}
2	HUMANE (Yoon et al., 2025)	29.19	0.1933 _{0.005}	0.4299 _{0.001}	0.2707 _{0.004}
3	Yellow Flash (Dharamvaram and Hakak, 2025)	31.71	0.1561 _{0.006}	0.4098 _{0.008}	0.2527 _{0.005}
4	FZIGOT (Rolinger and Liu, 2025)	18.50	0.3622 _{0.007}	0.3998 _{0.003}	0.2440 _{0.002}
5	EFC (Upravitelev et al., 2025)	7.01	0.1254 _{0.001}	0.3520 _{0.006}	0.2047 _{0.003}
6	Checkmate (Rashid and Hakak, 2025)	22.73	0.1848 _{0.007}	0.3368 _{0.005}	0.2043 _{0.005}
7	Baseline	33.88	0.2723 _{0.001}	0.3362 _{0.004}	0.2023 _{0.007}
8	OldJoe (Ftouhi et al., 2025)	48.57	0.1823 _{0.005}	0.3878 _{0.001}	0.1517 _{0.003}
–	CTU AIC (4o)	–	0.5035 _{0.003}	0.4373 _{0.004}	0.2690 _{0.004}
–	CTU AIC (4o-mini)	–	0.5718 _{0.005}	0.4809 _{0.003}	0.3176 _{0.001}

Table 2: Overall results for the AVERITeC shared task. Performance is evaluated on the total of 1000 hidden test set examples. Scores are given in Ev2R Recall for question-only, question-answer performance, and the total score.

quantization (Gerganov, 2023). With these efficiency modifications, five out of seven teams (HUMANE, Yellow Flash, FZIGOT, EFC, and Checkmate) achieved faster runtimes than the baseline’s average of 33.88s per claim.

The only non-baseline system that does not use model quantization is CTU AIC. Instead, Team CTU AIC uses the largest model with the maximum possible context size that fits on the VM’s GPU while satisfying the efficiency constraint, relying on the inherent processing abilities of the latest language models. While this results in the slowest runtime of all systems (53.67s average per claim), their system ranks highest in the shared task.

Particularly noteworthy is Team EFC’s runtime performance with an average of 7.01s per claim during inference, which is almost five times faster than the baseline. In addition to the aforementioned efficiency improvements, they proposed a semantic filtering step that reduces LLM calls by predicting the NEI or conflicting evidence/cherry-picking label using exclusively cosine similarity on retrieved evidence.

Despite the training cost not being considered in this shared task’s efficiency constraint, most teams did not train or fine-tune language models for any parts of their pipeline. The only exceptions are the systems of Team HUMANE and Team FZIGOT, discussed later in the report.

We further compare shared task systems to solutions using proprietary closed-source language models in Table 2. We modified the

winning system (CTU AIC) to use OpenAI’s GPT-4o (gpt-4o-2024-08-06) and GPT-4o-mini (gpt-4o-mini-2024-07-18) instead of Qwen3-14B. While question-only (Q only) scores increased substantially with closed models, both Q+A and AVeriTeC scores were lower than the original open-source CTU AIC system. Since we did not optimize the proprietary models for use in CTU’s system, these results provide only a preliminary assessment of their performance, as evidenced by GPT-4o-mini outperforming GPT-4o.

Question Generation Several teams (OldJoe, EFC, Yellow Flash, Checkmate, FZIGOT) begin claim verification by generating questions to guide evidence retrieval, following findings from the first shared task that question generation, rather than searching evidence for the claim directly, improves retrieval performance (Schlichtkrull et al., 2024). To generate the questions all teams rely on language models without further fine-tuning, specifically Qwen2.5, Qwen3, and Phi-4.

FZIGOT adopts an iterative question generation approach using a Graph-of-Thoughts framework (Besta et al., 2024). At each iteration, their system produces multiple questions, prunes similar ones, and verifies the claim using answers collected from these questions. If the label is "Not Enough Evidence" (NEE), the algorithm returns to question generation for a fixed number of iterations. FZIGOT uses LoRA (Hu et al., 2022) to fine-tune Qwen2.5-14B model for this step. Since the AVeriTeC training data is not structured in such iter-

Team Name	QG	Retrieval	QA	Veracity	Inference Engine	Efficiency
CTU AIC	Qwen3-14B	mxbai-embed-large-v1	Qwen3-14B	Qwen3-14B	Ollama	Dense Index
HUMANE	Qwen3-8B	gte-base-en-v1.5, Llama-3.1-8B, Qwen3-8B	Qwen3-8B	Qwen3-32B	vLLM	Dense Index, AWQ
Yellow Flash	Qwen2.5-7B	BM25, bilingual-embedding-small, snowflake-arctic-embed-m-v2.0	–	Phi-4-14B	vLLM	BM25, Chunking, GPTQ-int4
FZIGOT	Qwen2.5-14B	BM25, stella_en_400M	Qwen2.5-14B	Qwen2.5-14B	vLLM	Dense Index, LoRA, AWQ
EFC	Phi-4-14B	BM25, thenlper/gte-base	–	Phi-4-14B	llama.cpp	BM25, Chunking, Semantic Filtering, GGUF
Checkmate	Qwen2.5-7B	BM25, snowflake-arctic-embed-m-v2.0	–	Phi-4-14B	vLLM	BM25, Chunking, GPTQ-int4
OldJoe	Qwen3-14B	BM25, jina-embeddings-v3	Qwen3-14B	Qwen3-14B	vLLM	Dense Index, AWQ
Baseline	Llama-3.1-8B	BM25, SFR-embedding-2, Llama-3.1-8B	–	Llama-3.1-8B	vLLM	Retrieval cut-off

Table 3: Components used by shared task systems, ordered based on AVeriTeC-score (see Table 2). - indicates that the answer used was the entire retrieved passage.

ative fashion, FZIGOT creates a weakly supervised training dataset by generating a training instance for each question in the dataset, conditioning subsequent questions on previous questions accordingly. Their system achieves the highest Question-only EV2R Recall across teams with a score of 0.3622.

In contrast, CTU AIC and HUMANE produce questions by conditioning the generation on already retrieved evidence, following the baseline’s design. Team CTU AIC generates questions jointly with answers and the veracity prediction conditioned on retrieved evidence using Qwen3-14B. Since Team CTU AIC and HUMANE achieved the highest AVeriTeC scores, this suggests that relevant evidence can be retrieved from the provided knowledge store without explicit question generation. However, as described in Section 2.1, the knowledge store construction itself relies heavily on both annotators and models generating questions to find suitable evidence. As reported in the

AVeriTeC paper (Schlichtkrull et al., 2023a), search with generated questions yields complete evidence in 9/20 cases, compared to 16/20 with annotator-written questions. Using the same claims, we find that searching for *only* the claim yields complete evidence in 6/20 cases, whereas the full process of knowledge store construction (i.e., including the full list of queries described in Appendix A), complete evidence is found via search for 19/20 (for the shared task, the knowledge store is also extended with gold evidence, ensuring completeness also for the final claim). Since all systems use this provided knowledge store, question generation remains an integral part of every system. Additionally, all systems use generated questions and answers for veracity prediction, as discussed further below.

Evidence Retrieval Team EFC and Team FZIGOT retrieve evidence directly based on the generated questions. Team Yellow Flash and Checkmate additionally generate synthetic answers,

Team name	QV	N	E/P	C	PS	S	R	NEE	CE/C	Avg. # Docs
CTU AIC	0.49	0.17	0.40	0.41	0.46	0.18	0.4	0.1	0.06	9.0
HUMANE	0.30	0.19	0.37	0.33	0.35	0.27	0.35	0.0	0.0	10.0
Yellow Flash	0.19	0.16	0.27	0.33	0.35	0.23	0.26	0.02	0.06	7.27
FZIGOT	0.28	0.15	0.31	0.33	0.42	0.22	0.29	0.07	0.0	15.2
EFC	0.28	0.16	0.26	0.23	0.42	0.19	0.27	0.0	0.0	10.0
Checkmate	0.19	0.16	0.26	0.28	0.35	0.23	0.24	0.0	0.0	5.21
OldJoe	0.02	0.13	0.2	0.22	0.15	0.23	0.18	0.0	0.0	3.96
Average	0.25	0.16	0.3	0.3	0.36	0.22	0.28	0.03	0.02	8.66

Table 4: 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.

which are used to expand search queries for evidence retrieval. Team OldJoe formulates four distinct search queries for each question and retrieves evidence for each query individually. Team HUMANE applies a query expansion strategy that generates hypothetical fact-checking articles for each claim. This approach is also used by the baseline and their system from last year’s shared task. Team CTU AIC retrieves evidence by using the claim itself as the search query.

Similar to the first AVeriTeC shared task, teams explored vector-based dense retrieval systems (Karpukhin et al., 2020) and hybrid systems that combine dense retrieval with BM25 (Robertson and Zaragoza, 2009). Three systems (Team CTU AIC, FZIGOT, and HUMANE) relied solely on dense retrieval. Team HUMANE further summarizes the collected evidence into a single paragraph using Qwen3-8B. The remaining teams adopted hybrid retrieval approaches, following the baseline. Team Yellow Flash further groups together semantically similar sentences before embedding these coherent chunks and querying for dense retrieval.

Compared to fully dense retrieval, hybrid systems allow faster evidence retrieval by restricting neural search to a smaller subset of the knowledge store. This is reflected in the inference time reported by Team EFC. While Team OldJoe also employs a hybrid system, they create an index for both BM25 and dense embeddings over the entire knowledge store, and then combine retrieval scores using reciprocal rank fusion (Cormack et al., 2009).

Consistent with trends from the first shared task, models from the General Text Embeddings

(GTE) family (Li et al., 2023; Zhang et al., 2024) were widely adopted. These include Stella⁵ and the newer *snowflake-artic-embed-m-v2.0*, a GTE model fine-tuned using Matryoshka representation learning (Kusupati et al., 2022) to reduce quality degradation during model compression. Team CTU AIC used *mcbai-embed-large-v1* (Li and Li, 2024), the same retrieval model their team used in the previous shared task.

Question Answering & Veracity Prediction All teams used large language models for question answering and veracity prediction, relying on three models: Qwen3, Qwen2.5, and Phi-4. Three teams (Yellow Flash, EFC, and Checkmate) used retrieved evidence directly as answers, while Team HUMANE and OldJoe, who produce answers explicitly as a separate step in their verification pipeline, conditioned on claim, question, and evidence. Similarly to their question generation approach, Team FZIGOT uses LoRA to train distinct adapters for question answering and veracity prediction using weakly-supervised data. Apart from the increased efficiency during training, using three distinct adapters for each component of the pipeline can also improve inference runtime, as the loading and unloading of adapters into memory is substantially faster than for entire models. However, due to the experimental setting that allows systems to run one component of the pipeline at a time to account for restrictions of the VM, the effect of this design choice was less impactful in the context of the shared task.

⁵https://huggingface.co/dunzhang/stella_en_400M_v5

Team HUMANE submitted the only system with a fully fine-tuned model for veracity prediction. They trained a Qwen3-32B model and applied AWQ quantization to fit onto VM GPU memory. While Team CTU AIC did not fine-tune their model, they augmented the input with few-shot examples retrieved from the AVeriTeC training data, selected via BM25, conditioned on the claim. This shared task again highlights the importance of accurate veracity prediction components: top-ranking CTU AIC uses Qwen3-14B without quantization, while second-place HUMANE uses the largest language model (32B) with full-model fine-tuning.

Types & Verdicts Table 4 provides a detailed breakdown of results by claim type (quote verification, numerical claims, event/property claims, causal claims, and position statements) and verdict (supported, refuted, conflicting evidence/cherry-picking, and not enough evidence). We observe that all systems perform substantially worse on numerical claims compared to other claim types. While systems also underperformed on numerical claims in the first shared task, the performance gap is considerably larger this edition, which is likely contributed by the change in evaluation metric from Hungarian Meteor to EV2R.

Regarding performance across different veracity labels, no system achieves scores higher than 0.1 on Not Enough Evidence and Conflicting Evidence/Cherry-picking claims. This observation is expected and matches findings from the first shared task. These labels are highly challenging to correctly identify, subsequently causing some teams to omitting these labels from their predictions altogether. Moreover, systems that calibrate their veracity predictions to favor refuted claims gain an advantage (as long as they returned adequate evidence), as refuted claims dominate the dataset, comprising approximately two-thirds of all instances.

4 Human Evaluation of Evidence

Following the approach taken in last year’s AVeriTeC shared task (Schlichtkrull et al., 2024), 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 agree-

ment study of Schlichtkrull et al. (2023a). Second, the inaccuracies of automatic evaluation metrics of textual evaluation, require assessing and comparing the computed AVeriTeC scores with human annotations. 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. All seven teams were invited to participate in the evaluation. All teams but the team HUMANE took part in the evaluation. Each of the remaining six participating teams and two volunteers from with experience in automated fact-checking annotation manually evaluated 35 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 from and evenly distributed across all submitted systems, representing both high- and low-scoring systems, as shown in Table 4.

The figures in Appendix B show 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 evidence⁶. Annotators rated the evidence on a scale from 1 to 5 in two 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) **Relevance:** Measures how relevant the retrieved evidence is to the content of the claim.

Insights Gained The annotation process resulted in a total of 245 annotations. After filtering out evidence samples that were labeled by evaluators as not understandable (5 samples) or completely irrelevant to the given claim (11 samples), we were left with 229 valid annotations. Among these, 31 annotations corresponded to gold-labeled samples.

⁶We provide the exact instruction for rating each criteria in the appendix.

Label/Pred	CE/C	NEE	Refuted	Supported
CE/C	5.88	5.88	64.71	23.53
NEE	5.41	24.32	40.54	29.73
Refuted	3.96	5.94	77.23	12.87
Supported	5.88	1.96	11.76	80.39

Table 5: Overview of verdict **labelled** by human evaluators (rows) versus system **predictions** (columns) in percentages.

Excluding the gold-labeled samples, resulted in a final set of 198 evidence annotations.

Before labeling the system-retrieved evidence, participants were first asked to label the verdict given the retrieved evidence. Table 5 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 two evaluated dimensions (see Table 8), we find that the majority of predicted evidence was labeled as relevant (almost 80% evidence samples labelled as very relevant or mostly relevant to the claim), but in the dimension of semantic coverage, approximately 18% of the evidence received a rating of 1, indicating that “the predicted evidence covers none of the reference evidence.” Additionally, around 20% received a rating of 2, 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. Compared to the previous year’s AVeriTeC shared task, the relevance scores increased while the scores for semantic coverage remained roughly equal.

To assess the relationship between human scoring and the Ev²R score (see Sec 2.5), we computed both the Spearman correlation coefficient (ρ (Spearman, 1987)) and the Pearson correlation coefficient (r (Pearson, 1896)) as shown in Table 7. Correlations were calculated using both the entire evidence text and the question text only. In both cases, we observed a positive correlation between the AVeriTeC scores and the human evaluation (see Table 7) while the correlation with the coverage

Rating	COV	COV %	REL	REL %
1	35	17.68	2	1.01
2	47	23.74	9	4.55
3	40	20.20	30	15.15
4	45	22.73	84	42.42
5	31	15.66	73	36.87

Table 6: Overview of ratings for Semantic **Coverage** and **Relevance** scores obtained through human evaluation. Each score from 1 to 5 shows the absolute count and corresponding percentage.

Dimension	ρ	r
Coverage	.404	.406
Relevance	.244	.242

Table 7: Correlation between Q + A scores (AVeriTeC score) and human-rated subset of evidence. We calculate correlation using the Spearman (ρ) and Pearson (r) correlation coefficients.

dimension is higher than with relevance. Compared to last year’s shared task evaluation, where the correlation between manually assessed samples and the AVeriTeC score was close to zero for both coverage and relevance, this year’s score shows a much stronger alignment with human judgments (around 0.41 for coverage and 0.24 for relevance) when assessing the semantic coverage and relevance of predicted evidence. The human evaluation on the subset (see Table 8) shows a similar ranking of participating systems compared to automatic evaluations. 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.⁷ 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 10 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 word-

⁷See Table 8 in the appendix.

Team	Avg. Coverage	Leaderboard #
CTU AIC	3.6	1.
yellow flash	2.9	3.
HUMANE	2.9	2.
FZIGOT	2.9	4.
checkmate	2.1	6.
EFC	2.7	5.
OldJoe	2.4	7.

Table 8: Average semantic **coverage** scores assigned to evidence samples from selected teams based on human evaluation, next to AVeriTeC **rank** the team obtained in the 2025 shared task.

ing. Similarly, for the first two rows in Table 10, 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.

Certain low-ranked examples highlight different challenges (see Table 11). 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.

Acknowledgements

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). Michael is further supported by the Engineering and Physical Sciences Research Council (grant number EP/Y009800/1), through funding from Responsible AI UK (KP0016). The annotation of the new test set was conducted by a donation from Google.

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. Participating systems, which sought to maximize performance on AVeriTeC, may replicate those biases. While we constrained participants of using open-weights LLMs of a certain size, we did not enforce the use of open-data LLMs only, which would have been better in order to assess the biases in the participating systems. Open-weights models would also help to measure temporal leakage, as Qwen3, the most-used model in this shared task, has likely seen data that extended into the test set timeframe (January-December 2024), as it has an estimated training cutoff of March 2025. 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).

References

- Mubashara Akhtar, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2024. *Ev2r: Evaluating evidence retrieval in automated fact-checking*. *CoRR*, abs/2411.05375.
- 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.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoeffler. 2024. **Graph of thoughts: solving elaborate problems with large language models**. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.
- 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)*.
- Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. **Reciprocal rank fusion outperforms condorcet and individual rank learning methods**. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.
- Arbaaz Dharamvaram and Saqib Hakak. 2025. SANC-TUARY: An efficient evidence-based automated fact checking system. In *Proceedings of the Eighth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- 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.
- Elias Frantar, Saleh Ashkboos, Torsten Hoeffler, and Dan Alistarh. 2023. **OPTQ: Accurate quantization for generative pre-trained transformers**. In *The Eleventh International Conference on Learning Representations*.
- Farah Ftouhi, Russel Dsouza, Lance Gamboa, Jinlong Liu, Asim Abbas, Yue Feng, Mubashir Ali, Mark Lee, and Venelin Kovatchev. 2025. OldJoe at AVeriTeC: In-context learning for fact-checking. In *Proceedings of the Eighth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Georgi Gerganov. 2023. **ggml: Tensor library for machine learning**.
- 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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.
- 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*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. **Efficient memory management for large language model serving with pagedattention**. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.
- 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>.
- Xianming Li and Jing Li. 2024. **AoE: Angle-optimized embeddings for semantic textual similarity**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024.

- Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.
- Sewon Min, Kalpesh Krishna, Xinxu 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.
- 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.
- 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.
- 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.
- Farrukh Bin Rashid and Saqib Hakak. 2025. Fathom: A fast and modular RAG pipeline for fact-checking. In *Proceedings of the Eighth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Sascha Rolinger and Jin Liu. 2025. Graph-of-thoughts for fact-checking with large language models. In *Proceedings of the Eighth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Mark Rothermel, 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, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. 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.
- 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. 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.
- Herbert Ullrich and Jan Drchal. 2025. AIC CTU@FEVER 8: On-premise fact checking through long context RAG. In *Proceedings of the Eighth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.
- Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. 2025. Exploring semantic filtering heuristics for efficient claim verification. In *Proceedings of the Eighth 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.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers’ requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. [HerO at AVeriTeC: The herd of open large language models for verifying real-world claims](#). In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2025. Team HUMANE at AVeriTeC 2025: HerO 2 for efficient fact verification. In *Proceedings of the Eighth 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 9. 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-??, we include screenshots of the interface used by annotators. We also include, in Tables 10 and 11, instructive examples from the 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 9: 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:
 - **Supported**
 - **Refuted**
 - **Not Enough Evidence**
 - **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 "**predicted evidence**" (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 question below, you will only need to look at the **predicted evidence & claim!**

2. 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 6: 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 cyclists? 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 10: 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 11: Examples for system predictions which are low-ranked by the Averitec score.