Ev[2]R: Evaluating Evidence Retrieval in Automated Fact-Checking

Mubashara Akhtar[1,2], **Michael Schlichtkrull**[3] and **Andreas Vlachos**[4]

[1]Department of Computer Science, ETH Zurich, Switzerland [2]ETH AI Center, Switzerland
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[4]Department of Computer Science and Technology, University of Cambridge, UK
mubashara.akhtar@ai.ethz.ch, m.schlichtkrull@qmul.ac.uk
av308@cam.ac.uk

Abstract

Current automated fact-checking (AFC) approaches typically evaluate evidence either implicitly via the predicted verdicts or through exact matches with predefined closed knowledge sources, such as Wikipedia. However, these methods are limited due to their reliance on evaluation metrics originally designed for other purposes and constraints from closed knowledge sources. In this work, we introduce **Ev[2]R** which combines the strengths of reference-based evaluation and verdict-level proxy scoring. Ev[2]R jointly assesses how well the evidence aligns with the gold references and how reliably it supports the verdict, addressing the shortcomings of prior methods. We evaluate Ev[2]R against three types of evidence evaluation approaches: reference-based, proxy-reference, and reference-less baselines. Assessments against human ratings and adversarial tests demonstrate that Ev[2]R consistently outperforms existing scoring approaches in accuracy and robustness. It achieves stronger correlation with human judgments and greater robustness to adversarial perturbations, establishing it as a reliable metric for evidence evaluation in AFC.[1]

1 Introduction

To decide the truthfulness of a claim, professional fact-checkers search for, retrieve, and analyze evidence (Graves, 2018). Their goal is not only to assess a claim's accuracy, but also to present the evidence and fact-checking steps transparently, helping readers better understand and trust the fact-checking process (Graves, 2017).

Past research has primarily employed three approaches to evaluate evidence in Automated Fact-Checking (AFC). First, some methods assess evidence implicitly through the predicted verdict, assuming that accurate evidence leads to correct verdicts (Shahi and Nandini, 2020; Horne et al., 2018; Barrón-Cedeño et al., 2018). This overlooks the possibility that retrieved evidence may lead to a different verdict from the gold evidence without directly assessing the evidence itself. Second, some approaches restrict evidence retrieval to predefined, closed knowledge sources like Wikipedia, comparing the retrieved evidence to reference evidence from the same source and considering only exact matches correct. While common in AFC benchmarks (Thorne et al., 2018a; Jiang et al., 2020; Sathe et al., 2020; Nørregaard and Derczynski, 2021), this method may dismiss valid evidence not annotated as part of the reference, ignoring the broader variety of trusted open-web sources, such as government statistics. Lastly, in recent datasets requiring evidence retrieval from the entire internet (Schlichtkrull et al., 2023; Chen et al., 2024), evaluation often uses token-matching metrics like METEOR (Banerjee and Lavie, 2005). However, these metrics are highly sensitive to surface form differences and fail to account for alternative 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 *liters.''* can both be valid steps on the way to establish relative levels of alcohol consumption between South Africa and other countries.

In this paper, we introduce Ev[2]R, a weighted scorer for evidence evaluation in AFC that combines two complementary scoring signals: (i) a reference-based scoring component, which assesses retrieved evidence through comparison with reference evidence data, and (ii) a second

[1]Code is available at `https://github.com/mubasharaak/fc-evidence-evaluation`.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 530–561, 2026. https://doi.org/10.1162/TACL.a.647
Action Editor: Wei Gao. Submission batch: 4/2025; Revision batch: 7/2025; Published 4/2026.

component, which evaluates evidence using the annotated verdict label as proxy reference. The reference-based scoring enhances interpretability by clearly demonstrating evidence accuracy (precision) and completeness (recall) relative to annotations. Moreover, for a given claim, multiple sets of evidence or alternative reasoning paths may lead to the same (correct) verdict. As such, looking at verdicts can give (noisy) signal for evidence evaluation. Hence, we incorporate verdict-level alignment via a proxy-reference scoring component into Ev^{2}R. Taking both into account, the Ev^{2}R score reflects both the factual correctness and completeness of the retrieved evidence with respect to the reference, as well as its alignment with the verdict.

We evaluate Ev^{2}R across three AFC datasets (i.e., AVeriTeC [Schlichtkrull et al., 2024], FEVER [Thorne et al., 2018a], and VitaminC [Schuster et al., 2021]) and compare it against evidence scorers used in previous research, including reference-based, proxy-reference, and reference-less baselines. AVeriTeC contains real-world claims with web-based evidence, reflecting practical challenges in fact-checking. FEVER and VitaminC are widely used large-scale benchmarks, each providing multiple evidence sets per claim, sometimes with conflicting verdicts (e.g., one supporting and another refuting the same claim). Together, these datasets allow us to evaluate Ev^{2}R across diverse settings.

We assess Ev^{2}R based on its correlation with human ratings and its robustness to adversarial tests with perturbed evidence. Our results demonstrate that the Ev^{2}R scorer outperforms baselines in terms of correlation with human ratings across all datasets. Additionally, adversarial stress tests reveal that Ev^{2}R is more robust to common perturbations, such as redundancy and noise, compared to traditional metrics used in previous research. These findings highlight that combining reference- and verdict-level scoring provides a more reliable approach for evidence evaluation in AFC.

2 Related Work

2.1 Evidence-based AFC

While earlier AFC datasets evaluate evidence implicitly through the predicted veracity label (Guo et al., 2022; Akhtar et al., 2023), more recent approaches require explicit evaluation to ensure that system predictions are grounded in evidence.

There are several challenges in the automated retrieval and evaluation of evidence for evidence-based AFC. First, evidence-based AFC datasets released in recent years contain purpose-built claims derived from Wikipedia, such as FEVER (Thorne et al., 2018a), Hover (Jiang et al., 2020), WikiFactCheck (Sathe et al., 2020), and ChartCheck (Akhtar et al., 2024). These datasets restrict potential evidence to Wikipedia pages from which the claims were derived. While this restriction is unrealistic, it allows the use of simple, conditional scoring functions for evidence evaluation. For example, the FEVER-Score (Thorne et al., 2018b) assess the verdict conditioned on retrieving the correct evidence.

To create a more realistic setting, recent datasets use evidence retrieved from the entire internet, such as AVeriTeC (Schlichtkrull et al., 2023) and MultiFC (Augenstein et al., 2019). While their claims and evidence setup better align with real-world fact-checking, this complicates evaluation. Retrieved evidence can no longer be evaluated using exact matches on document identifiers or paragraphs. Second, for a given claim, there may be more than one correct set of evidence pieces, as different reasoning chains with varying evidence can be used to verify the claim. For example, to contradict the claim, ''Dr. Fauci stated during a recent press statement that wearing masks in outdoor spaces is not recommended,'' evidence can state that the speaker was not Dr. Fauci or prove that the content of the statement itself is incorrect.

Finally, correct evidence may be wrongly interpreted as incorrect if it became available after the dataset's creation and is therefore not included in the set of annotated evidence within the dataset.

2.2 Evaluation Approaches for Natural Language Generation

To evaluate evidence in the context of AFC, related work on evaluating natural language generation (NLG) systems can serve as inspiration due to the similarly open-ended nature of the task, e.g., a system can return multiple equally valid summaries for the same input text which should be evaluated against a single gold reference summary (Celikyilmaz et al., 2020). While earlier NLG evaluation methods used string-matching metrics to compare generated and reference text based on heuristics (Celikyilmaz et al., 2020; Lin, 2004; Papineni et al., 2002), more recent approaches, involve training language models for evaluation (Zhang et al., 2020; Sellam et al., 2020) or use large language models' (LLMs') conditional language generation capabilities (Fu et al., 2023).

Reference-based Evaluation. Reference-based metrics evaluate the generated text $\hat{x}$ by comparing it to a gold reference x:

$$y = f(x, \hat{x}),$$

where the scoring function $f(.)$ can be conceptualized in various ways, for example, as a non-parametric function based on n-grams or a similarity function based on conceptualized representations of text. Earlier NLG metrics used n-grams to compare generated and reference text based on different heuristics such as string overlap, string distance or lexical diversity (Celikyilmaz et al., 2020; Lin, 2004; Papineni et al., 2002). In recent years, various LLM-based metrics were proposed, which also serve as inspiration for Ev^{2}R. BLEURT, NUBIA, BERTScore, and MoverScore are all evaluation metrics that use contextualized representations (such as BERT (Devlin et al., 2019) embeddings) to assess the similarity between generated and reference text. BLEURT is a learned metric trained on synthetic sentence pairs (Sai et al., 2023), NUBIA (Kane et al., 2020) follows a three-step approach for similarity scoring, BERTScore uses token-level similarity with greedy matching (Zhang et al., 2020), and MoverScore allows partial matching between n-grams/words for more flexible evaluation (Zhao et al., 2019).

Proxy-reference Evaluation. This family of metrics uses a proxy for evaluation and does not directly compare the generated text ($\hat{x}$) to the annotated reference text (x). In previous research they have been applied if the gold references were not available for evaluation or to focus on specific evaluation criteria (e.g., factuality) for which appropriate proxies are defined as we highlight below (Nema and Khapra, 2018; Durmus et al., 2020; Huang and Zhang, 2021). Multiple proxy-based metrics for NLG evaluation were proposed in the past (Eyal et al., 2019; Scialom et al., 2019). Previous works use proxy references like questions and answers, relational tuples, or entailment labels to evaluate generated text. For example, Durmus et al. (2020) assess if relevant information from the input text is present in the generated text, while other works use relational tuples to check if the generated text matches factual information from the reference text (Goodrich et al., 2019). Additionally, entailment labels are used to detect hallucinations in generated text, based on the idea that a summary should entail its source document (Ji et al., 2023).

Reference-less Evaluation. This family of metrics evaluates generated text without the availability of any (proxy) reference. More recently, the capabilities of autoregressive LLMs (e.g., GPT4 [OpenAI, 2023]) have been used for reference-less evaluation. Fu et al. (2023) define evaluation as a text generation problem. Their proposed GPTScore framework is based on the idea that LLMs are more likely, i.e., with higher conditional generation probability, to produce the generated text (e.g., a summary) if it is of high quality and fulfills given specifications based on the input document. FactScore (Min et al., 2023) is a recent LLM-based framework for reference-less evaluation of factuality in LLM-generated text. To evaluate the factuality of generated text, it breaks down the text into atomic facts and verifies if they are supported by the knowledge sources. The overall score corresponds to the ratio of supported facts.

3 Ev^{2}R: A Weighted Scorer for Evidence Evaluation

This section introduces Ev^{2}R, a weighted scorer for evaluating evidence in the context of AFC. The scorer combines two distinct but complementary parts: a reference-based component, which decomposes the comparison between retrieved and reference evidence into precision and recall (see Section 3.1); and a proxy-based component, which assesses alignment between the predicted and reference verdict labels, where the verdict label acts as a proxy for evaluating evidence accuracy (see Section 3.2). In Section 3.3, we provide an overview of the score calculation, integrating both components.

3.1 Reference-based

The reference-based scorer $\mathcal{S}_{\text{ref-based}}$ compares the retrieved and reference evidence by decomposing them into atomic facts (more details below) and

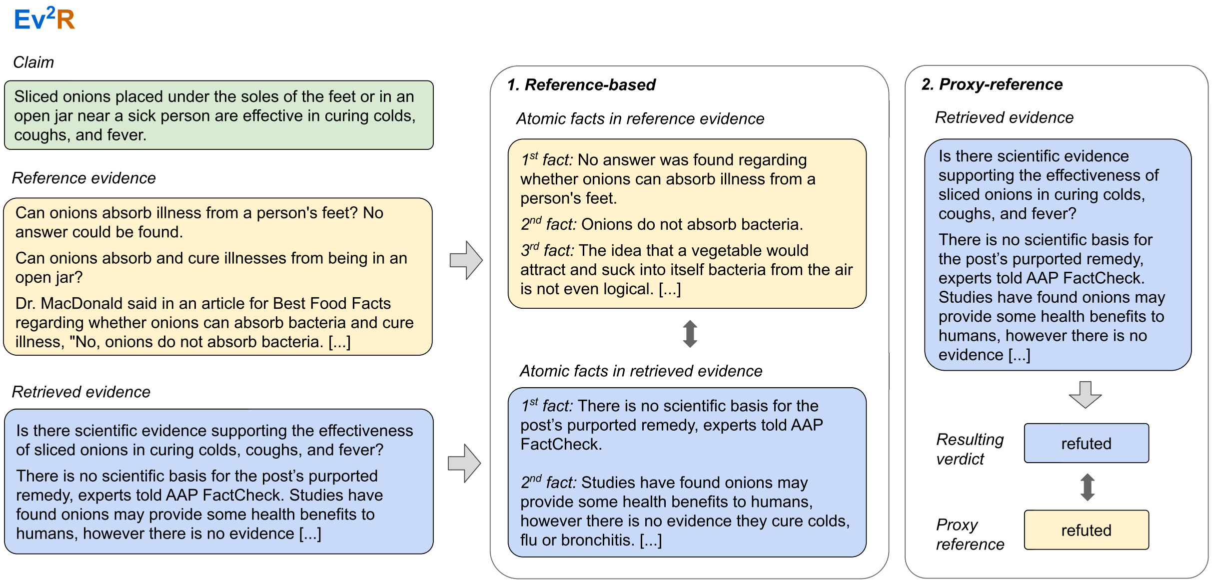


Figure 1: Overview of the proposed **Ev²R** scorer, including the reference-based (middle) and proxy-reference (right) components. For reference-based evaluation, evidence is first decomposed into atomic facts before evaluation. Proxy-reference scoring uses the verdict label as proxy to assess retrieved evidence.

evaluating their alignment using a precision and a recall score:

$$S_{\text{ref-based}}(\hat{E}, E) = (s_{\text{prec}}, s_{\text{recall}})$$

The precision score s_{prec} measures how accurately the retrieved atomic facts align with the reference, while s_{recall} evaluates the completeness of the retrieved facts compared to the reference. Inspired by Min et al. (2023), our scorer first splits the retrieved evidence $\hat{E}$ and reference evidence E into atomic facts, $A_{\hat{E}}$ and A_E, respectively (see the middle box in Figure 1). In Figure 1 we demonstrate the extraction of atomic facts from evidence. We decompose the following evidence *'Is there scientific evidence supporting the effectiveness of sliced onions in curing colds, coughs, and fever? There is no scientific basis for the post's purported remedy, experts told AAP FactCheck. Studies have found onions may provide some health benefits to humans, however there is no evidence [...]'* into two facts, first, *'There is no scientific basis for the post's purported remedy, experts told AAP FactCheck.'*, and second, *'Studies have found onions may provide some health benefits to humans, however there is no evidence they cure colds, flu or bronchitis. [...]'*

The reference-based evaluation component assesses the factual consistency of a predicted evidence by systematically comparing it to the reference evidence.[2] First, the scorer breaks down both the predicted and reference evidence into atomic facts, where each fact is a separate sentence. Then, each fact from the predicted evidence is evaluated to verify whether it is directly supported by any fact in the reference evidence, and vice versa, i.e., each reference fact is checked for support in the predicted evidence. We instruct the scorer to rely on the provided texts, disallowing any external sources or background knowledge. The final output summarizes the total number of facts in each evidence (i.e., reference and retrieved) and how many of them are supported, offering a measure of factual alignment.

Based on the extracted and scored facts, the scorer computes precision and recall scores to evaluate evidence quality: Precision reflects how many retrieved evidence facts are supported by the reference evidence, while recall measures how many reference evidence facts are round in the retrieved evidence. We specify the precision score s_{prec} as the ratio of facts supported by the reference evidence:

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

[2] Figure 4 in the Appendix shows the exact prompt we use for extracting and verifying facts.

The scorer iterates over each fact ($a_{\hat{E}} \in A_{\hat{E}}$), applying an indicator function ($\mathbb{I}[a_{\hat{E}}$ supported by $E]$) that returns 1 if the fact $a_{\hat{E}}$ is supported by the reference evidence E and 0 otherwise. Similarly, the recall score captures whether each atomic fact of the reference evidence ($a_E \in A_E$) is supported by the retrieved evidence, evaluating the extent to which the retrieved evidence covers the reference evidence:

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

In addition to precision and recall, we also provide an aggregated F_1 score (s_{f_1}). While the reference-based scorer enables fine-grained reference-based evaluation, it may underestimate retrieved evidence when it is correct but different from the reference set. The proxy component, we introduce in Section 3.2, helps stabilize evaluation in such cases by assessing the stance of the retrieved evidence against the reference verdict.

We evaluate the reference-based component with several LLMs as backbone models: GPT4o, Gemini 1.5 Pro, Gemini 1.5 Flash, and Llama 3.1 70B. We provide further details on the models in Section 4 and the instruction prompts in the Appendix (see Figures 2, 3, and 4).

3.2 Proxy-reference

The proxy-reference scoring component of Ev^{2}R uses the verdict label y as a proxy to assess the retrieved evidence $\hat{E}$. As shown in the right box of Figure 1, this component compares the predicted verdict label to the reference label, i.e., *refuted*. Proxy-reference scoring uses a fine-tuned DeBERTa language model (He et al., 2021) as backbone to predict the verdict label y serving as a proxy for evidence evaluation.

The proxy-reference scoring component builds on a DeBERTa-v3 model (He et al., 2021) fine-tuned to predict verdict labels for claim-evidence pairs. We initialized the scorer from a checkpoint pretrained on a diverse set of NLI datasets, i.e., MNLI (Williams et al., 2018), FEVER-NLI (Nie et al., 2019), Adversarial-NLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), and WANLI (Liu et al., 2022), which provide a foundation for assessing factual inference. To adapt the model for evidence evaluation in the context of automated fact-checking, we fine-tuned it further on AFC datasets, including the training sets of FEVER, VitaminC, HOVER, and AVeriTeC. The training was conducted using the HuggingFace Trainer API, mixed-precision (fp16) and an initial learning rate of 1e^{-6}. The model was trained for three epochs, a warmup phase of fifty steps, and evaluated every $10k$ steps using the micro-F1 score on held-out evidence sets to select the best checkpoint.

Given a claim c, a retrieved evidence set $\hat{E}$, and a reference label $y \in \mathcal{Y}$, the proxy scoring component $\mathcal{S}_{\text{proxy}}$ assigns a confidence score s_{proxy} to the label y, reflecting the model's confidence in the predicted verdict:

$$\begin{aligned} s_{\text{proxy}} &= \mathcal{S}_{\text{proxy}}(c, y, \hat{E}) \\ &= \text{softmax}(\mathbf{z})_y = \frac{e^{z_y}}{\sum_{y' \in \mathcal{Y}} e^{z_{y'}}} \end{aligned}$$

where $\mathbf{z}$ is the vector of logits over all possible veracity labels $\mathcal{Y}$. The score $s_{\text{proxy}} \in [0, 1]$ is computed as the softmax of the logits $\mathbf{z}$, where the index y corresponds to the reference label, and the softmax function transforms the logits into a probability distribution over the possible labels. The resulting softmax distribution quantifies the model's confidence.

3.3 Final Weighted Scorer

To compute the final Ev^{2}R score, we use the outputs of both previously introduced components: the reference-based scorer $\mathcal{S}\text{ref}(\hat{E}, E)$, which yields precision and recall scores ($s$prec and s_{recall}), and the proxy scorer $\mathcal{S}_{\text{proxy}}(c, y, \hat{E})$, which estimates the confidence in the label y. We first compute the F_1 score from the precision and recall outputs of $\mathcal{S}_{\text{ref}}$. Then we calculate the final weighted score as a combination of the resulting F_1 score and the s_{proxy}. Specifically, we use a weighting factor $\alpha = 0.5$ to equally balance both components:

$$s_{\text{Ev}^2\text{R}} = \alpha \cdot \left(\frac{2 \cdot s_{\text{prec}} \cdot s_{\text{recall}}}{s_{\text{prec}} + s_{\text{recall}}} \right) + (1 - \alpha) \cdot s_{\text{proxy}}$$

This ensures that the final score reflects both the factual correctness and completeness of the retrieved evidence with respect to the reference, as well as its alignment with the verdict.

4 Evaluation of Ev^{2}R

This section provides an overview of the systematic evaluation approaches conducted as part of the meta-evaluation of Ev^{2}R, which involves three different data sources. First, using adversarial tests generated through perturbing evidence data (see Section 4.1). Second, evidence obtained from AFC systems submitted to the AVeriTeC shared task competition (Schlichtkrull et al., 2024) in early 2024. (see Section 4.2). Third, evidence pairs extracted from the AFC benchmarks FEVER (Thorne et al., 2018a) and VitaminC (Schuster et al., 2021). We discuss the data sources in more detail below. Finally, in Section 4.4, we give an overview of baseline scorers and models used for evaluation.

4.1 Evaluation with Adversarial Tests

We evaluate the scorers by assessing their robustness against adversarial perturbations, including both semantics-altering and semantics-preserving changes. Specifically, we perform adversarial stress tests, including both semantics-altering and semantics-preserving perturbations, such as redundancy, noise insertion, and completeness alterations. Inspired by Ribeiro et al. (2020), we generated large-scale test sets by perturbing either the claim or evidence data. We included both semantics-preserving and semantics-modifying tests (see Table 6 in the Appendix). For semantics-preserving tests, we change the claim and/or evidence such that the resulting evaluation score should not be affected, for example, by introducing small typos. These tests assess the scorers' robustness to semantically equivalent changes in the evidence data. On the other hand, we also create semantics-modifying tests by changing the evidence such that its meaning alters and subsequently scores should drop, for example, by removing a significant part of the evidence. To create tests with automated perturbations, we use the AVeriTeC test set as the basis.

4.2 Evaluation with Human Ratings

Data Source The AVeriTeC dataset involves retrieving evidence from the web and assessing the veracity of real-world claims previously checked by professional fact-checkers. It addresses limitations in prior AFC benchmarks through focusing on real-world claims and annotated evidence in question-answer format. The AVeriTeC shared task attracted a total of 21 submissions of fact-checking systems (Schlichtkrull et al., 2024). Participating systems were required to provide URLs for their retrieved evidence as well as verdict labels categorized as supported, refuted, not enough evidence, or conflicting evidence/cherrypicking. The organizers provided a public knowledge store with documents to avoid reliance on costly search APIs.

In collaboration with the AVeriTeC organizers, we obtained a subset of evidence across all submitted systems, equally representing both high- and low-scoring systems. After ranking the systems according to their performance on the shared task, we categorize them in four performance groups each containing approximately 25% of submitted systems to allow comparative analysis: low score (below 0.16), mid-low score (between 0.16 and 0.22), mid-high score (between 0.22 and 0.435), and high score (0.435 or higher). This classification enabled systematic examination of system performance across different scoring levels. Using stratified sampling, we selected 560 claims to ensure an even distribution across the dataset.

Human Ratings and Agreement The shared-task participants were then asked to evaluate these samples.[3] Thirteen participating teams manually annotated thirty evidence samples each, five of which were gold-labeled samples annotated by ourselves. For the final set, we only considered annotation submissions of a team, if the gold-labeled samples were correctly annotated, resulting in a final set consisting of 278 evidence annotations. To measure the agreement between annotators, we used two methods. First, we calculated the agreement between the categorical verdict labels that annotators selected as the first step in the annotation process. Therefore, we used all 278 annotated samples after filtering as specified above, and obtained a Krippendorff's Alpha (K-α) (Krippendorff, 2004) 0.732 and a Fleiss' Kappa (Fleiss, 1971) (F-κ) of 0.727—both indicating substantial agreement. Second, for the rating dimensions which are annotated with numeric values between one and five, we calculated the standard deviation (std) among the rating, resulting in std between 0.804 and 1.318 (see Table 1).

Evaluation Dimensions With the help of human raters, we assessed how well the scorers correlate with human judgments. For this evaluation, we considered three dimensions:

(1) **Coverage** assesses how much of the reference evidence is covered by the retrieved

[3]Find details on the annotation platform in the Appendix.

Rating Dimension	IAA method	Score
Verdict agreement	Fleiss-κ	0.727
Verdict agreement	Krippendorff-α	0.732
Coverage	Std	1.119
Relevance	Std	0.804

Table 1: Inter-annotator agreement (IAA) scores for human ratings using Fleiss-κ/Krippendorff-α for categorical (i.e., verdict agreement) and standard deviation (Std) for continuous ratings.

evidence, whether the content, meaning, entities, etc., of the reference evidence are fully represented in the retrieved evidence, ensuring that the retrieved evidence covers all aspects of the reference evidence necessary for the evaluation. Partial evidence may overlook certain details and result in incomplete or incorrect conclusions.

(2) **Relevance** measures the relevance of the evidence retrieved for the claim, ensuring that the evaluated evidence addresses the claim and that minimal unrelated information is included.

(3) **Verdict Agreement** measures if the retrieved evidence results in the same verdict label as the reference evidence. Discrepancies in verdicts can indicate issues with the retrieved evidence such that a follow-up review can be performed to ensure accuracy in the fact-checking process.

While coverage can penalize potentially valid evidence, this only occurs when such evidence is included at the expense of the gold/reference evidence. As long as the gold evidence is fully covered, the coverage score remains high, even if additional content is included. This differs from the relevance score we introduced above, which is more flexible: It does not penalize the inclusion of content different to the gold evidence, as long as the content is relevant for verifying the claim. We report both alignment with relevance and coverage, with the latter being relevant for the recall-oriented automatic Ev^{2}R evaluation.

4.3 Evaluation with Additional Benchmarks

In addition to the AVeriTeC shared task data, we use the large-scale AFC benchmarks FEVER (Thorne et al., 2018a) and VitaminC (Schuster et al., 2021). FEVER is a benchmark designed for automated claim verification against textual sources from Wikipedia. It contains $185.4k$ claims generated through modifying sentences from Wikipedia and classified as supported/refuted/not enough information. VitaminC is an AFC benchmark designed for robust claim verification. It consists of approximately $400k$ pairs of claims and evidence pairs derived from Wikipedia revisions. The aim was to create contrastive pairs of nearly identical evidence, differing only by subtle factual details, such that one evidence set supports the claim and the other refutes it, or vice versa. For scorer assessment, we created tuples consisting of a claim c_i, a reference evidence set E_{ref}, predicted evidence set E_{pred} from the multiple evidence sets available for individual claims within the FEVER and VitaminC test datasets. If both evidence sets resulted in the same verdict label, we assigned a verdict agreement score of 1; otherwise, we assigned a score of 0. The resulting data were used for evaluations presented in Table 2.

4.4 Baselines

To validate the effectiveness of Ev^{2}R, we compare it to three types of baseline scorers: (i) the reference-based BLEURT scorer, which compares retrieved evidence to reference evidence; (ii) proxy-reference scorers, which assesses evidence through predicted verdict labels; and (iii) reference-less scorers, which rely only on the claim and retrieved evidence.[4]

(i) **Reference-based Baseline Scorers** As a reference-based baseline, we adapt a scorer built on BLEURT (Sellam et al., 2020), originally developed for NLG tasks such as summarization. BLEURT uses BERT as its backbone and is trained on millions of synthetically generated sentence pairs to capture semantic similarities between texts. To evaluate the retrieved evidence against reference evidence, the BLEURT-based scorer compares both evidence sets and gives a similarity score. The final aggregated score s is calculated by averaging instance-level scores across all evaluated samples.

Given a set of retrieved evidence $\hat{E} = \{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n\}$ and reference evidence $E = \{e_1, e_2, \ldots, e_n\}$, the scorer computes:

$$s = \text{BLEURT}(E, \hat{E})$$

[4]Prompts and further details are given in the Appendix.

Scorer	VitaminC		FEVER		AVeriTeC		Avg	
	ρ	r	ρ	r	ρ	r	$\lvert\rho\rvert$	$\lvert r\rvert$
Reference-based Baselines								
BLEURT	$-.065$	$-.072$	$-.003$	$-.000$	.151	.098	.073	.057
RougeL	.030	.022	$-.010$	$-.009$	.070	.097	.037	.043
BLEU	.030	.026	$-.006$	$-.003$	.166	.137	.067	.055
Meteor	.030	.021	$-.010$	$-.003$	.152	.159	.064	.061
H-METEOR	.001	$-.001$	$-.027$	$-.016$	$-.029$	$-.019$	.019	.012
Reference-less Baselines								
GPT4o	.014	.015	.457	.468	.240	.243	.237	.242
Gemini-Pro	.034	.033	.415	.412	.262	.273	.237	.239
Gemini-Flash	.038	.045	.434	.426	.230	.247	.234	.239
Llama 3.1	.024	.034	.293	.308	.215	.208	.177	.183
Ev^{2}R Weighted Scorer								
GPT-4o / Proxy-ref	.276	.318	.366	.360	.500	.488	.381	.389
Gemini-Pro / Proxy-ref	.295	.334	.349	.346	.472	.464	.372	.381
Gemini-Flash / Proxy-ref	.336	.346	.170	.240	.129	.197	.292	.294
Llama / Proxy-ref	.272	.322	.026	.227	.324	.307	.243	.273

Table 2: Correlation on the **Verdict Agree** category calculated using the Spearman (ρ) and Pearson (r) correlation coefficients for the datasets VitaminC, FEVER, and AVeriTeC (shared task subset). Averages are computed over the absolute values of each metric per row. Highest scores per column are colored blue and second-highest values highlighted with brown color.

where $s \in [0, 1]$, with higher values indicating greater semantic similarity between the evidence.

We fine-tune BLEURT for evidence evaluation using the PyTorch version of the original BLEURT model[5] and the HuggingFace Trainer API over 5 epochs, with a batch size of 4 and learning rate set to 1e^{-5}. The model is trained on claim-evidence input pairs from AFC datasets including FEVER (Thorne et al., 2018a), VitaminC (Schuster et al., 2021), and HOVER (Jiang et al., 2020). These datasets provide multiple annotated evidence sets $E_1, E_2, \ldots, E_n$ per claim, enabling the construction of diverse similarity pairs. We generate input-label pairs where positive examples (similarity score $s = 1$) are created using two distinct evidence sets aligned to the same claim or based on matching claim-verdict pairs (e.g., both supporting the claim). Negative examples ($s = 0$) are created by pairing evidence from different claims or by contrasting verdict-aligned pairs (e.g., one supporting, one refuting).

Additionally, we include metrics such as RougeL, BLEU, METEOR, and Hungarian METEOR as reference-based baselines since they have been previously also used in AFC for evidence evaluation (Schlichtkrull et al., 2024). They were originally developed for NLG tasks and compare retrieved ($\hat{E}$) with reference evidence (E) based on surface level or semantic similarity.

(ii) **Proxy-reference LLM Baseline** As a proxy-reference baseline, we propose an LLM-based scorer that estimates the likelihood of the correct veracity label y given the claim c and the retrieved evidence $\hat{E}$. Instead of relying on a predicted probability distribution as in trained classifiers, we prompt the LLM directly with an instruction to predict the veracity label and use its log-probability for the correct label as the proxy score. The scorer computes a score s as follows:

$$s = \log p(y \mid P(I, c, \hat{E}), \theta)$$

where $P(I, c, \hat{E})$ is the prompt including an instruction I, the input claim c, and the retrieved evidence $\hat{E}$, and θ denotes the parameters of the LLM. We apply chain-of-thought prompting (Wei et al., 2022) and instruct the model to generate intermediate reasoning steps before selecting a final label. The prompt includes few-shot examples and

[5]Elron/bleurt-base-512.

Scorer	VitaminC		FEVER		AVeriTeC		Avg	
	ρ	r	ρ	r	ρ	r	$\|\rho\|$	$\|r\|$
Reference-based (Precision)								
GPT4o	.203	.200	**.140**	**.129**	**.278**	**.256**	**.207**	**.195**
Gemini-Pro	**.222**	**.216**	.139	.118	.237	.177	.199	.170
Gemini-Flash	.202	.198	.106	.079	−.290	−.205	.199	.161
Llama 3.1	−.037	−.036	.110	.109	.216	.210	.121	.118
Reference-based (Recall)								
GPT4o	.229	.228	.146	.114	**.314**	**.284**	.230	.209
Gemini-Pro	**.253**	**.246**	**.176**	**.143**	.285	.278	**.238**	**.222**
Gemini-Flash	.232	.227	.032	.000	−.190	−.039	.151	.089
Llama 3.1	.020	.025	−.011	−.012	.128	.121	.053	.053
Reference-based (F_1)								
GPT4o	.215	.213	.143	.121	**.295**	**.269**	**.218**	**.202**
Gemini-Pro	**.237**	**.230**	**.155**	**.129**	.259	.216	.217	.193
Gemini-Flash	.216	.212	.049	.000	−.230	−.066	.172	.115
Llama 3.1	.087	.164	−.024	−.027	.161	.154	.074	.073
Proxy-reference								
Proxy-ref	.456	.479	.290	.480	.487	.459	.411	.473

Table 3: More detailed scores for both Ev^{2}R components: the reference-based part (using different LLMs as backbone models) and the DeBERTa-based proxy-reference component.

reasoning steps as demonstrations to improve prediction quality. For the proxy-reference scorer, we use DeBERTa as the backbone model. Our choice is motivated by its strong performance on claim verification and natural language inference (NLI) tasks under fine-tuned settings, as well as its compatibility with our available compute resources (one A100 GPU).

(iii) **Reference-less Baseline** As a reference-less baseline, we instruct an LLM-based scorer to evaluate the retrieved evidence $\hat{E}$ based solely on the input claim c, without relying on any reference evidence. The scorer decomposes the claim into atomic facts A_c and checks whether each fact $a_c \in A_c$ is addressed, for example, supported or refuted, by the retrieved evidence. The score is computed as follows:

$$s = \frac{1}{|A_c|} \sum_{a_c \in A_c} \mathbb{I}[a_c \text{ supported/refuted by } \hat{E}]$$

This approach, similar to the Ev^{2}R reference-based component, is inspired by FactScore (Min et al., 2023). However, unlike the reference-based evaluation, which extracts evidence for both gold and retrieved evidence, this method decomposes the claim to evaluate the retrieved evidence without relying on any additional reference.[6]

Backbone Models. For the LLM-based scorers, we evaluate several backbone models: GPT4o, Gemini 1.5 Pro, Gemini 1.5 Flash, and Llama 3.1 70B (see Table 2). GPT4o is a state-of-the-art model that demonstrates strong performance across a wide range of benchmarks and domains (OpenAI, 2024). We also evaluate Gemini 1.5 Pro, which has shown strong performance in reasoning and long-context understanding tasks (Reid et al., 2024), while Gemini 1.5 Flash, a smaller variant, allows us to explore scalability across different model sizes. We also include Llama 3.1 70B (Dubey et al., 2024) to understand how well our proposed scorers perform with state-of-the-art open source models compared to the previously mentioned closed-source models.

5 Results and Discussion

In this section, we discuss the results obtained with Ev^{2}R using three different data sources (see Tables 2 and 3). First, human-rated system predictions from the AVeriTeC shared task (see Table 4).

[6]The reference-less evaluation prompt is given in Figure 2 in the Appendix.

	Ref-less Baselines				Proxy-ref	Ref-based Baselines				Ev²R			
	GPT-4o	Gem-Pro	Gem-Flash	Llama	DeBERTa	RougeL	BLEU	Meteor	H-Met	GPT-4o	Gem-Pro	Gem-Flash	Llama
COV (ρ)	.237	.287	.275	.297	.338	.150	.236	.229	.005	.321	.341	.221	.328
COV (r)	.261	.296	.287	.286	.348	.169	.184	.240	$-.024$	.323	.326	.299	.336
REL (ρ)	.292	.203	.287	.227	.298	.086	.107	.062	.008	.297	.278	.403	.224
REL (r)	.360	.263	.333	.250	.374	.099	.079	.076	.003	.332	.315	.404	.277

Table 4: Correlation between human-rated **AVeriTeC shared task submissions** and proposed scorers for COVerage and RELevance dimensions. Highest scores per dimension are colored blue and second-highest values highlighted with brown color.

Dataset	Reference-less Baselines					Ev²R			Proxy-ref
	METEOR	ROUGE	BLEU	BLEURT	GPT4o	Prec	Rec	Proxy-comp.	DeBERTa
Semantics-altering tests									
Completeness	-30.56	-53.19	-77.87	**-143.70**	**-91.76**	-21.2	-59.5	-39.35	-82.68
Random shuffle	-32.32	-89.94	**-94.20**	-1.02	0.55	-6.8	-6.0	-32.22	**-92.32**
Average	-31.44	-71.57	**-86.04**	-72.36	-45.61	-14.0	-32.75	-35.79	**-87.50**
Semantics-preserving tests									
Invariance contraction	-0.53	-0.97	-1.35	-1.65	**0**	-0.8	-0.7	-0.85	**0.02**
Invariance num2text	-8.34	-10.24	-14.10	-22.13	-3.85	**0.1**	**0**	-4.02	-35.64
Invariance text2num	-0.04	-0.90	-1.15	-0.81	-4.40	**0**	-0.3	-1.59	**0.01**
Invariance synonyms	**-8.22**	-36.64	-51.83	-51.00	**-2.75**	-11.6	-11.3	-15.35	-85.71
Redundancy sent	-26.16	-46.52	-56.15	-80.45	3.30	**0**	**-0.1**	-5.04	-67.95
Redundancy words	-5.73	-10.68	-32.69	-31.90	-2.75	**-1.0**	**0.2**	-2.89	-52.55
Fluency	-11.52	-23.33	-29.26	-41.69	**-2.20**	-4.3	-3.9	**-2.6**	-57.12
Noise	-31.82	-31.81	-31.81	-21.20	**-1.65**	-28.7	-2.2	**2.89**	-62.42
Argument structure	-0.01	**0**	**0**	**0**	**1.65**	**0**	**0**	-0.74	-28.93
Average	-10.33	-18.43	-23.94	-27.87	-15.04	-5.16	**-2.9**	**-3.78**	-39.79

Table 5: Results obtained for adversarial tests as score difference (in %) between initial evidence and manipulated evidence. The evaluated Ev²R reference-based scoring uses $GPT4o$. Overall, the reference-based scoring and the proxy-reference scoring components demonstrate the most robust performance on adversarial tests, together with the trained reference-based baselines, as evidenced by small performance drops on semantics-preserving tests. Best scores per evaluation category are colored blue and second-best values highlighted with brown color.

Second, the FEVER and VitaminC test sets for which we obtained pairs of evidence following the approach outlined in Section 4.4. Third, adversarial tests as described in Table 5. Following Fu et al. (2023), we measure the correlation between the scorers and human ratings using Spearman (ρ) (Spearman, 1987) and Pearson correlation coefficients (r) (Pearson, 1896).

General Insights Across Datasets. Overall, the trained proxy-reference scorer yields the highest overall correlation across the datasets we assess, except for AVeriTeC (see Table 2). Unlike the other benchmarks, which contain claims and evidence extracted from Wikipedia (Thorne et al., 2018a; Schuster et al., 2021), AVeriTeC contains more complex, real-world claims extracted from fact-checking websites and evidence from the web. Hence, the trained proxy-reference scorer is outperformed by the more recent LLM GPT4o on AVeriTeC. Moreover, across almost all datasets in Table 2 and all rating categories in Table 4, our Ev²R scorers rank among the top-2 scorers. The obtained insights support the practical usefulness of the Ev²R scorer for evaluating evidence in realistic AFC scenarios.

Further analysis across all datasets demonstrates that the evaluated Ev²R scorer components result in statistically significant correlations across backbone models selected for reference-based scoring. We assessed statistical significance using two-sided hypothesis tests on the correlation coefficients, as implemented in `scipy.stats.pearsonr` and `scipy.stats.spearmanr`, with the null hypothesis that no correlation (i.e., $r = 0$ or $\rho = 0$) is present. We conducted all tests using $n = 278$ data points from the AVeriTeC dataset for which we had collected human ratings as specified in Section 4.2, as well as the FEVER and VitaminC test sets preprocessed as specified in Section 4.3. The proxy reference scoring component achieves the strongest correlations across datasets ($\rho = 0.375$, $r = 0.393$ for FEVER; $\rho = 0.250$, $r = 0.273$ for VitaminC; $\rho = 0.475$, $r = 0.450$ for AVeriTeC; all p-values < 0.001). Among the other models, GPT-4o and Gemini-Pro generally demonstrate good predictive power (with all p-values < 0.001), while Gemini-Flash results in weaker, but still statistically significant correlations. Our results show variability in models for reference scoring across benchmarks.[7]

Ev^{2}R Weighted Score versus Traditional Metrics. Across all datasets, i.e., VitaminC, FEVER, and AVeriTeC, the Ev^{2}R scorer consistently outperforms traditional reference-based baselines we assess, including ROUGE-L, BLEU, METEOR, and H-METEOR. The performance gap is particularly prevalent on AVeriTeC, which includes real-world claims and evidence retrieved from the web. This setting is notably more complex than earlier benchmarks like FEVER and VitaminC. In such cases, traditional metrics that rely on surface-level or shallow semantic similarity often fail to generalize. These findings are supported by the insights we gain from the adversarial tests in Table 5, where we assess all baselines against both both Ev^{2}R components, the proxy reference (last column) and reference-based (prec/recall) components. On all semantics-preserving tests, both Ev^{2}R components yield robustness performance.

Ev^{2}R: Balancing Proxy and Reference-based Evidence Evaluation. While the proxy-reference component of Ev^{2}R helps stabilize scores on academic datasets with short claims and evidence (e.g., FEVER, VitaminC in Table 2), relying solely on proxy signals proves insufficient for more complex real-world scenarios such as AVeriTeC (see Table 2). Prior work (McCoy et al., 2019) has shown that NLI models often exploit shallow syntactic heuristics such as lexical overlap or subsequence matching that can lead to correct predictions for the wrong reasons. This motivated the need for a balanced evaluation approach that combines proxy-reference with reference-based assessment and offers a more reliable and interpretable measure of evidence data. Moreover, for relevant assessment dimensions such as *Coverage* of reference evidence and *Relevance* of retrieved evidence, the reference-based component of Ev^{2}R yields higher correlation with human assessment.

Selection of Backbone LLMs. Assessing the impact of different backbone LLMs for the reference-based component of Ev^{2}R, we find that the model selection has only a limited effect on the overall performance of the Ev^{2}R scorer across datasets (Table 2) and evaluation dimensions (Tables 2 and 4). No single model consistently outperforms others across all dimensions and datasets. While scorer variants using Gemini models correlate higher with human assessments in the dimensions coverage and relevance, the GPT-4o-based variant excels in the category of verdict agreement (Table 2). Moreover, the combination of GPT-4o and proxy-reference shows strong correlations on AVeriTeC and is competitive overall.

Adversarial Tests. In Table 5, we assess the performance of Ev^{2}R components (reference-based and proxy-reference) across various adversarial tests against baselines, considering both semantics-preserving and semantics-altering tests. Our insights highlight how Ev^{2}R components (i.e., reference-based and proxy-reference) versus baseline scorers respond to adversarial manipulations in evidence texts. Traditional metrics such as METEOR, ROUGE, BLEU, and especially BLEURT show high sensitivity to semantics-altering tests, demonstrated through substantial performance drops. On the other hand, the Ev^{2}R components, specifically proxy-reference scoring, exhibit stronger robustness to both semantics-altering and semantics-preserving evaluation. These findings highlight their suitability and robustness for evidence evaluation. Overall, Ev^{2}R demonstrates greater robustness compared to baselines in adversarial settings.

5.1 Qualitative Analysis of Evidence Evaluation

Reference-based. Reference evidence provides additional context to the Ev^{2}R scorer, which can be leveraged during the evaluation of retrieved evidence. The reference-based scoring component

[7] See Table 9 for detailed correlation results.

of Ev^{2}R allows a more detailed and interpretable evaluation by breaking down both reference and predicted evidence into atomic facts (see Table 10 in the Appendix). Reporting separate precision and recall scores facilitates a more detailed understanding of the retrieved evidence. For instance, retrieved evidence sometimes contains excessive details compared to the reference, negatively affecting precision but still yielding high recall.

While this approach enhances the evaluation process, our qualitative analysis identified challenges that are addressed by stabilizing the Ev^{2}R scorer through integration of a proxy-reference component. One specific challenge for reference-based scoring are cases where predicted evidence uses different information or follows a different reasoning chain than the reference evidence, but results in the same verdict (e.g., refuting a claim, as illustrated in Table 10 in the Appendix). In such cases, the reference-based scoring tends to yield lower scores compared to reference-less and proxy-reference scorers. For example, in Table 10 both reference and predicted evidence contradict the claim that Nigeria is the fifth largest recipient of diaspora remittances. However, the predicted evidence states that Nigeria is the sixth largest recipient, while the reference states it is the seventh largest. Hence, the scorer identifies a mismatch between the reference and predicted evidence. Furthermore, retrieved evidence can sometimes include more details than the reference, which impacts precision but leads to a high recall. To address these limitations and stabilize the Ev^{2}R scorer, we incorporate proxy-reference scoring.

Proxy Reference. We also observe that a key challenge for proxy-reference scoring are evidence pairs where both predicted and reference evidence are related but result in diverging verdict labels (see Table 12 in the Appendix). For example, considering the claim *''Scotland is spending more on health per head than the rest of the UK.''* in Table 12; while the predicted evidence closely resembles the reference in surface-level characteristics, the predicted evidence supports the claim, whereas the reference evidence does not, resulting in a low proxy-reference score. Similarly, minor numerical differences, such as those observed in evidence in Table 12, can also lead to diverging verdicts despite overall consistency between the reference and predicted evidence. In these cases, a more interpretable assessment based on precision and recall scores from reference-based scoring can yield further insights and enhance readers' understanding of the presented evidence.

5.2 How Ev^{2}R Addresses Key Challenges of Evidence Evaluation for AFC

This section reiterates key challenges of evidence evaluation (see Sections 1 and 2) and Ev^{2}R.

Implicit Evidence Assessment. Verdict-based evaluation approaches can inaccurately rate evidence high by predicting correct verdicts without genuinely relying on the retrieved evidence. To overcome this over-reliance on verdict alone, Ev^{2}R explicitly evaluates both alignment with reference evidence and verdict agreement. Evaluation results from AVeriTeC, FEVER, and VitaminC show high correlation with human judgments on criteria such as verdict agreement, evidence relevance, and coverage of reference data, indicating Ev^{2}R assesses retrieved evidence rather than verdict correctness alone.

Assessment Relying on Closed Knowledge Sources & Surface-level Matching of Evidence. Previous widely used approaches for evidence evaluation (e.g., the FEVER score [Thorne et al., 2018a]) relied on predefined sources (e.g., Wikipedia) and exact matches between predicted and reference evidence. Such methods are not extendable to realistic AFC settings, where predicted evidence may not be part of the annotated corpus. Additionally, token-matching metrics (e.g., h-METEOR used in Schlichtkrull et al. [2024]) are overly sensitive to surface-level differences and fail to capture alternative valid evidence or reasoning paths. Ev^{2}R's reference-based component decomposes retrieved evidence into atomic facts and assesses their alignment with reference facts beyond strict matching constraints. Evaluations of evidence predicted with shared task systems (see Table 4), which include unannotated web evidence, confirm the effectiveness of this approach. Furthermore, adversarial test results (see Table 5) demonstrate Ev^{2}R's superior robustness against surface-level perturbations compared to baseline scorers.

Multiple Valid Evidence Sets and Reasoning Paths & Temporal Constraints with Evidence Annotation. For a given claim, multiple correct sets of evidence or alternative reasoning paths can exist. Previous evaluation methods might penalize valid alternative evidence. Ev^{2}R with its two scoring components considers scenarios with multiple valid reasoning paths and evidence absent in the annotated references. Evaluations using the datasets FEVER, VitaminC, and AVeriTeC, where exact matching between reference and predicted evidence is not expected, show Ev^{2}R achieves higher correlations with human judgments and stable performance under adversarial testing.

6 Conclusion

This paper introduces Ev^{2}R, a weighted evidence evaluation scorer for AFC. Ev^{2}R consists of a reference-based and a verdict-level proxy evaluation component and addresses limitations in previous evaluation approaches by jointly assessing how well evidence aligns with gold references and how reliably the evidence supports the verdict. We assessed Ev^{2}R against three baseline approaches for evidence evaluation approaches and find that Ev^{2}R correlates strongly with human judgments on relevant evaluation criteria and is a robust scorer for evidence evaluation in AFC.

7 Limitations & Ethics Statement

This work has several limitations that are important to acknowledge. First, the datasets used are restricted to English, limiting the generalizability of findings to global fact-checking contexts where multilingual capabilities are essential. This English-only focus reflects a broader issue in the field, where most benchmarks are available solely in English, potentially distorting perceptions of progress in automated fact-checking. Additionally, all labels, e.g., supports and refutes used in this work to classify claims, capture the relationship between claims and evidence, they do not imply any real-world correctness of the claims. Finally, we informed the participants about the data being collected and its purpose. Participants had the opportunity to withdraw at any time and to provide feedback.

Acknowledgments

We thank Nedjma Ousidhoum, Julian Eisenschlos, Oana Cocarascu, and Elena Simperl for their valuable feedback and suggestions about this work. Mubashara Akhtar acknowledges funding from the ETH AI Center Postdoctoral Fellowship. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958). Michael Schlichtkrull is supported by the EPSRC grant AdSolve (EP/Y009800/1), a RAI UK Keystone project.

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.361

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.828

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. https://doi.org/10.18653/v1/D19-1475

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.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims. task 2: Factuality.In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.196

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783. https://doi.org/10.48550/arXiv.2407.21783

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.454

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1395

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among several raters. *Psychological Bulletin*, 76(5):378–382. https://doi.org/10.1037/h0031619

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166. https://doi.org/10.48550/ARXIV.2302.04166

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019*, pages 166–175. ACM. https://doi.org/10.1145/3292500.3330955

L. Graves. 2018. Understanding the promise and limits of automated fact-checking.

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture and Critique*, 10(3):518–537. https://doi.org/10.1111/cccr.12163

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net.

Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25–28, 2018*, pages 518–527. AAAI Press. https://doi.org/10.1609/icwsm.v12i1.14982

Nannan Huang and Xiuzhen Zhang. 2021. Evaluation of review summaries via question-answering. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 87–96, Online. Australasian Language Technology Association.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38. https://doi.org/10.1145/3571730

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.309

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.508

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

Sewon Min, Kalpesh Krishna, Xinxi 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. https://doi.org/10.18653/v1/2023.emnlp-main.741

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1429

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*. https://doi.org/10.1609/aaai.v33i01.33016859

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: Claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774. https://doi.org/10.48550/ARXIV.2303.08774

GPT4o OpenAI. 2024. Blog post: Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-10-23.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.421

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. https://doi.org/10.1098/rsta.1896.0007

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530. https://doi.org/10.48550/arXiv.2403.05530

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.442

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2):26:1–26:39. https://doi.org/10.1145/3485766

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.

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. https://doi.org/10.18653/v1/2024.fever-1.1

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the Web. *CoRR*, abs/2305.13117. https://doi.org/10.48550/arXiv.2305.13117

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. https://doi.org/10.18653/v1/2021.naacl-main.52

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1320

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid- A multilingual cross-domain fact check news dataset for COVID-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020 Workshops, Atlanta, Georgia, USA [virtual], June 8, 2020*.

C. Spearman. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471. https://doi.org/10.2307/1422689, PubMed: 3322052

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. https://doi.org/10.18653/v1/N18-1074

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. https://doi.org/10.18653/v1/W18-5501

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing*

Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28–December 9, 2022.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1053

A Evaluation with Adversarial Perturbations

#	Test Type	Description
1	Completeness	Drops substantial parts of the evidence, causing the veracity label to change, i.e., $y \in \{\text{support; refute}\}$ to *not enough information (NEI)*. We expect Ev^{2}R scorers to react to these changes with a reduction in scores.
2	Random Shuffle	Randomly shuffles all words occurring in an evidence such that the resulting evidence text is not understandable.
3	Fluency	Semantics-preserving tests that (a) introduce typos or (b) drop stop words.
4	Invariance	Introducing invariant changes to the evidence data: (a) changing numbers to text (e.g., 2 to *two*); (b) changing numbers as text to numerals (e.g., *two* to 2); (c) replacing words in evidence by their synonyms; (d) introducing contractions (e.g., *we're* instead of *we are*).
5	Noise	Adds noise by inserting a random sentence from another evidence set in the dataset.
6	Redundancy	Introduces redundancy by duplicating sentences or words within sentences.
7	Argument Structure	Changes the order of evidence sentences to disrupt the logical flow of arguments.

Table 6: Categories of adversarial tests with descriptions.

B Additional Evaluation Dimensions

As an additional data source for the evaluation of scorers, we collected ratings for 100 randomly selected test data instances from the AVeriTeC test set (Schlichtkrull et al., 2023), together with the evidence retrieved through the AVeriTeC baseline system as described in Schlichtkrull et al. (2023). Each test sample was evaluated by two annotators, who were either computer science graduate students or postdoctoral researchers familiar with AFC research. This resulted in 200 annotations while the overall rating for each annotated sample and dimension was calculated as the average of both annotations. If the annotators disagreed on the verdict assigned to the retrieved evidence, hence interpreted the relationship between retrieved evidence and claims differently, we assigned the data instance to a third annotator and used majority voting to determine the final verdict.
Table 8 given an overview over correlation between scorers and human ratings on three additional dimensions, shortly outlined below.

(1) **Coherence** captures whether the retrieved evidence is coherent, i.e., all sentences are connected sensibly and the evidence makes sense as a whole. An important aspect of fact-checking is presenting the verification process to the reader in a way that is comprehensible. Hence, well-structured and coherent evidence supports understanding of the verification process and the rationale behind the resulting verdict.

(2) **Repetition** evaluates whether the retrieved evidence exhibits repetition of its content. As mentioned earlier, supporting readers' understanding of the fact-checking process can be enhanced through avoiding repetitive content in the evidence. This can further maintain the reader's focus on unique information in the evidence text, leading to better understanding of the resulting conclusion, i.e., the verdict.

(3) **Consistency** assesses whether the retrieved evidence is logically consistent in itself, since consistency in evidence improves the reliability of the fact-checking process and trust among readers.

Scorer	Coverage		Coherence		Repetition		Consistency		Relevance	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
Prompt Scorer GPT4o										
Ref-less	.237	.261	.201	.254	−.006	−.009	.112	.118	.292	.360
Ref-based (prec)	.277	.270	.185	.191	−.089	−.090	.218	.242	.316	.313
Ref-based (recall)	.339	.334	.199	.197	−.140	−.105	.187	.190	.277	.269
Ref-based (F_1)	.306	.298	.192	.194	−.109	−.097	.201	.213	.295	.289
Proxy-ref	.253	.306	.142	.201	.007	−.011	.133	.140	.182	.221
Prompt Scorer Gemini-Pro										
Ref-less	.287	.296	.208	.246	.012	.010	.092	.110	.203	.263
Ref-based (prec)	.304	.254	.186	.163	−.094	−.115	.193	.197	.270	.264
Ref-based (recall)	.398	.380	.282	.268	−.168	−.164	.199	.205	.248	.248
Ref-based (F_1)	.345	.304	.224	.203	−.120	−.135	.196	.201	.259	.256
Prompt Scorer Gemini-Flash										
Ref-less	.275	.287	.256	.269	.027	.019	.147	.146	.287	.333
Ref-based (prec)	.086	.232	.105	−.004	−.481	−.450	.337	.181	.490	.368
Ref-based (recall)	.132	.270	.103	.106	−.577	−.619	.330	.291	.527	.531
Ref-based (F_1)	.104	.249	.104	−.008	−.524	−.521	.333	.223	.508	.435
Prompt Scorer Llama 3.1										
Ref-less	.297	.286	.222	.210	−.011	−.008	.134	.116	.227	.250
Ref-based (prec)	.341	.336	.178	.181	−.120	−.122	.080	.111	.131	.172
Ref-based (recall)	.299	.313	.174	.179	−.060	−.048	.148	.143	.175	.187
Ref-based (F_1)	.318	.324	.176	.180	−.080	−.069	.104	.125	.150	.179
Trained Scorer										
Ref-based	.116	.033	.091	.063	−.133	.024	.057	.059	.145	.110
Proxy-ref	.338	.348	.230	.286	−.011	−.057	.293	.329	.298	.374
Baselines										
RougeL	.150	.169	.180	.190	.057	.040	.124	.131	.086	.099
BLEU	.236	.184	.180	.166	−.144	−.039	.040	.038	.107	.079
Meteor	.229	.240	.192	.191	−.150	−.132	.061	.064	.062	.076
H-METEOR	.005	−.024	.076	.057	.117	.025	.039	.024	.008	.003
Weighted Scorer ($\alpha = 0.5$)										
GPT-4o / Proxy-ref	.321	.323	.211	.240	−.060	−.077	.247	.271	.297	.332
Gemini-Pro / Proxy-ref	.341	.326	.227	.244	−.066	−.096	.245	.265	.278	.315
Gemini-Flash / Proxy-ref	.221	.299	.167	.147	−.267	−.289	.313	.276	.403	.404
Llama 3.1 / Proxy-ref	.328	.336	.203	.233	−.046	−.063	.199	.227	.224	.277

Table 7: Correlation between human-rated **AVeriTeC shared task submissions** and proposed scorers, across five quality dimensions. Highest scores per column are colored blue and second-highest values highlighted with brown color. Reference-based precision and recall scores which are calculated to compute the final reference-based F_1 score, are colored in gray.

Scorer	Coverage		Coherence		Repetition		Consistency		Relevance	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
Prompt Scorer GPT4o										
Ref-less	.237	.261	.201	.254	$-.006$	$-.009$	.112	.118	.292	.360
Ref-based (prec)	.277	.270	.185	.191	$-.089$	$-.090$	.218	.242	.316	.313
Ref-based (recall)	.339	.334	.199	.197	$-.140$	$-.105$	.187	.190	.277	.269
Ref-based (F_1)	.306	.298	.192	.194	$-.109$	$-.097$	.201	.213	.295	.289
Proxy-ref	.253	.306	.142	.201	.007	$-.011$	.133	.140	.182	.221
Prompt Scorer Gemini-Pro										
Ref-less	.287	.296	.208	.246	.012	.010	.092	.110	.203	.263
Ref-based (prec)	.304	.254	.186	.163	$-.094$	$-.115$	.193	.197	.270	.264
Ref-based (recall)	.398	.380	.282	.268	$-.168$	$-.164$	.199	.205	.248	.248
Ref-based (F_1)	.345	.304	.224	.203	$-.120$	$-.135$	.196	.201	.259	.256
Prompt Scorer Gemini-Flash										
Ref-less	.275	.287	.256	.269	.027	.019	.147	.146	.287	.333
Ref-based (prec)	.086	.232	.105	$-.004$	$-.481$	$-.450$	.337	.181	.490	.368
Ref-based (recall)	.132	.270	.103	.106	$-.577$	$-.619$	.330	.291	.527	.531
Ref-based (F_1)	.104	.249	.104	$-.008$	$-.524$	$-.521$	.333	.223	.508	.435
Prompt Scorer Llama 3.1										
Ref-less	.297	.286	.222	.210	$-.011$	$-.008$	.134	.116	.227	.250
Ref-based (prec)	.341	.336	.178	.181	$-.120$	$-.122$	.080	.111	.131	.172
Ref-based (recall)	.299	.313	.174	.179	$-.060$	$-.048$	.148	.143	.175	.187
Ref-based (F_1)	.318	.324	.176	.180	$-.080$	$-.069$	.104	.125	.150	.179
Trained Scorer										
Ref-based	.116	.033	.091	.063	$-.133$	.024	.057	.059	.145	.110
Proxy-ref	.338	.348	.230	.286	$-.011$	$-.057$	.293	.329	.298	.374
Baselines										
RougeL	.150	.169	.180	.190	.057	.040	.124	.131	.086	.099
BLEU	.236	.184	.180	.166	$-.144$	$-.039$	.040	.038	.107	.079
Meteor	.229	.240	.192	.191	$-.150$	$-.132$	.061	.064	.062	.076
H-METEOR	.005	$-.024$	.076	.057	.117	.025	.039	.024	.008	.003
Weighted Scorer ($\alpha = 0.5$)										
GPT-4o / Proxy-ref	.321	.323	.211	.240	$-.060$	$-.077$	.247	.271	.297	.332
Gemini-Pro / Proxy-ref	.341	.326	.227	.244	$-.066$	$-.096$	.245	.265	.278	.315
Gemini-Flash / Proxy-ref	.221	.299	.167	.147	$-.267$	$-.289$	.313	.276	.403	.404
Llama 3.1 / Proxy-ref	.328	.336	.203	.233	$-.046$	$-.063$	.199	.227	.224	.277

Table 8: Correlation between human-rated **AVeriTeC shared task submissions** and proposed scorers, across five quality dimensions. Highest scores per column are colored blue and second-highest values highlighted with brown color. Reference-based precision and recall scores which are calculated to compute the final reference-based F_1 score, are colored in gray.

Scorer Component	Spearman ρ		Pearson r	
	Correlation	p-value	Correlation	p-value
FEVER				
Gemini Flash Precision	0.139	0.0018	0.119	0.0080
Gemini Flash Recall	0.139	0.0019	0.105	0.0193
Gemini Pro Precision	0.139	0.0019	0.118	0.0082
Gemini Pro Recall	0.176	7.86e-05	0.143	0.0013
GPT-4o Precision	0.140	0.0016	0.129	0.0038
GPT-4o Recall	0.146	0.0011	0.114	0.0107
Proxy Score	0.375	3.48e-18	0.393	6.29e-20
VitaminC				
Gemini Flash Precision	0.202	5.10e-06	0.198	8.49e-06
Gemini Flash Recall	0.232	1.48e-07	0.227	2.87e-07
Gemini Pro Precision	0.222	5.27e-07	0.216	1.14e-06
Gemini Pro Recall	0.253	9.65e-09	0.246	2.66e-08
GPT-4o Precision	0.203	4.53e-06	0.200	6.67e-06
GPT-4o Recall	0.229	2.37e-07	0.228	2.70e-07
Proxy Score	0.250	1.53e-08	0.273	5.63e-10
AVeriTeC				
Gemini Flash Precision	−0.081	0.803	−0.034	0.00916
Gemini Flash Recall	−0.051	0.875	0.030	0.00925
Gemini Pro Precision	0.272	1.42e-05	0.226	0.00034
Gemini Pro Recall	0.282	6.97e-06	0.269	1.86e-05
GPT-4o Precision	0.283	1.68e-06	0.260	1.17e-05
GPT-4o Recall	0.316	7.31e-08	0.284	1.47e-06
Proxy Score	0.475	1.29e-15	0.450	5.99e-14

Table 9: Correlation between reference and predicted scores for the evaluation dimension ‘verdict agreement’ on FEVER, VitaminC, and AVeriTeC.

C Evaluation Examples

Id	Claim	Reference Evidence	Predicted evidence
602	Nigeria is the fifth largest receiver of diaspora remittances in the world.	*Which countries are the top 5 largest receivers of diaspora remittances in the world?* *In 2020, the top 5 countries were India, China, Mexico, the Philippines and Egypt.* *Where does Nigeria rank in the list of countries that are the largest receivers of diaspora remittances?* *Nigeria remains the seventh-largest recipient.* (verdict: refuted)	*What is the definition of diaspora remittances?* *Diaspora remittances are money transfers made by citizens living outside their home country.* *What is the ranking of Nigeria in terms of diaspora remittances?* *Nigeria is ranked as the sixth largest receiver of remittances in the world.* (verdict: refuted)

Table 10: Example from a system submitted to AVeriTeC shared task. Both reference and predicted evidence contradict the claim but represent alternative evidence paths.

Id	Claim	Reference evidence	Predicted evidence
602	Global warming is occurring because of solar winds and changes in Earth's magnetic field.	*Does solar wind affect global warming?* *Most of the particles from solar wind is reflected back into space by the Earth's magnetic field. Because the energy does not reach our planet's surface, it has no measurable influence on surface temperature.* *Do changes in earth's magnetic field impact global warming?* *There's no known physical mechanism capable of connecting weather conditions at Earth's surface with electromagnetic currents in space.* *What is causing global warming?* *Scientists attribute the global warming trend observed since the mid-20th century to the human expansion of the ''''greenhouse effect''''—warming that results when the atmosphere traps heat radiating from Earth toward space.*	*What causes global warming?* *Global warming is primarily caused by human activity, specifically the emission of carbon dioxide (CO2) and other greenhouse gases. These gases trap excess heat on Earth, leading to an increase in the planet's average temperature.* *What is the role of solar winds in global warming?* *Solar winds, consisting of charged particles emitted by the sun, do not contribute to global warming. Earth's magnetic field prevents most of these particles from interacting with Earth. Changes in the sun's behavior cannot explain modern global warming. The warming trend is caused by human-emitted greenhouse gases trapping excess heat on the Earth.* *Can changes in Earth's magnetic field cause global warming?* *No, changes in Earth's magnetic field do not have an impact on global temperatures.* *What is the scientific consensus on the cause of global warming?* *The scientific consensus on the cause of global warming is that it is primarily caused by human activities, particularly the burning of fossil fuels which leads to the emission of greenhouse gases into the Earth's atmosphere. [...]*

Table 11: Example for predicted evidence with repetitions.

Id	Claim	Reference evidence	Predicted evidence
56	Scotland is spending more on health per head than the rest of the UK.	*How much does the UK spend on healthcare? The UK as a whole spends the equivalent of one quarter of total public expenditure and £2,972 per capita in 2017.* *How much does Scotland spend on healthcare? Scotland spends 21 percent or £8bn on healthcare, per person £2,507.* *How much does Northern Ireland spend on healthcare? 22 percent or £5bn and £2,616 per person.* *How much does Wales spend on healthcare? Wales spends 23 percent or £13.7bn and £2,546 per person.* *How much does England spend on healthcare? 25 percent or £136.6bn and £2,427 per person.*	*What is the health spending per head in Scotland? In 2013/14, Scotland spent £2,187 per head in real terms.* *What is the health spending per head in England? In 2013/14, England spent £2,028 per head in real terms.* *How does the health spending per head in Scotland compare to the rest of the UK? Scotland continues to spend the most per head on health compared to the rest of the UK.* *What is the trend in health spending per head in the UK since 2003/04? All four countries of the UK spend more per head on health now than they did in 2003/04, but the trend has flattened since 2009/10.* *What is the health spending per head in Northern Ireland? In 2013/14, Northern Ireland spent £2,158 per head in real terms.* *What is the health spending per head in Wales? In 2013/14, Wales spent £2,026 per head in real terms. [...]*

Table 12: Example from systems submitted to AVeriTeC shared task (Schlichtkrull et al., 2024) extracted from manual evaluation of proxy-reference evaluation.

Id	Claim	Facts	Predicted evidence
20	France24 encouraging Niger Delta militants to fight for their right to equity in resource sharing.	*1. France24 is encouraging Niger Delta militants.* *2. The encouragement is to fight for their right to equity in resource sharing.*	*Since groups in the Niger Delta could not be mobilising simply for the sake of doing so, the insight that these analysts attempt to proffer is that the Niger Delta mobilisation must be for the maximisation of sectional interests, with the non-producing ethnic groups a target of their grievance.*
593	Antifa vandalized the home of Missouri senator Josh Hawley in 4th of January 2021	*1. Antifa vandalized the home of Missouri senator Josh Hawley.* *2. The vandalism occurred on January 4, 2021.*	*Was there any evidence linking Antifa to the vandalism of Missouri senator Josh Hawley's home on January 4, 2021? In Newark, New Jersey, police and FBI investigators had identified ''a probable 'Antifa' related individual,'' who was arrested for possessing a knife, a hatchet, and a jar of gasoline.* *Where was Senator Josh Hawley residing at the time of the alleged Antifa vandalism on January 4, 2021? who was living in Omaha. Antifa Nebraska printed thousands. [...]*
648	Biden won fewer counties than Trump and Obama but more votes, suggesting election fraud in 2020.	*1. Biden won fewer counties than Trump.* *2. Biden won fewer counties than Obama.* *3. Biden won more votes than Trump.* *4. Biden won more votes than Obama.* *5. This suggests election fraud in 2020.*	*Did the increase in votes for Biden suggest any irregularities in the 2020 election? The claim that Biden won fewer counties than Trump and Obama but more votes doesn't necessarily suggest election fraud.It is possible for a candidate to win more votes by appealing to a broader base of voters in specific key areas, regardless of the number of counties won. Such an analysis requires a comprehensive understanding of the electoral process, including voter turnout, population distribution, and voting patterns. Without specific evidence of fraud, the claim lacks substantial support.*

Table 13: Examples from systems submitted to AVeriTeC shared task. The second column depicts *Facts* that are have been extracted from the respective claim for reference-less evidence evaluation.

D Prompt Templates

Reference-Less Prompt

You will get as input a claim and evidence.

Please verify the correctness of the claim following the following steps.
1. Break down the claim in independent facts. Each fact should be a separate sentence.
2. Only break down the claim into facts, not the evidence!
3. Evaluate each fact individually using the given evidence only. Do not use additional sources or background knowledge. Explain the evaluation.
4. Finally summarise how many facts are (1.) supported by the evidence, (2.) clearly are contradicted by the evidence, (3.) how many facts you have in total.

Generate the output in form of a json as shown in the example below.

Examples:

Claim: Mukesh Ambani, richest man in Asia had surgery for pancreatic cancer at Sloan Kettering, New York, US cancer speciality hospital on October 30, 2020.

Evidence: When was the photograph taken of Mukesh Ambani on the Facebook post claiming he had been diagnosed with pancreatic cancer and had undergone surgery? The photograph was taken on September 5, 2020. When was a video filmed of Mukesh Ambani at the virtual launch of NK Singh's book Portrait of Power? The video was filmed on October 19, 2020. What date was the Facebook post which confirmed Mukesh Ambani had lost 30 kgs, been diagnosed with pancreatic cancer and had had liver transplant surgery? The Facebook post was dated November 2, 2020. Where was Mukesh's photo of him supposedly recieving surgery actually taken? It was taken by Manushree Vijayvergiya who shared her experience of meeting Mukesh and Isha Ambani in a cafe in Liechtenstein.

Output: {{
"facts": "1. Mukesh Ambani is the richest man in Asia. 2. Mukesh Ambani had surgery for pancreatic cancer. 3. The surgery took place at Sloan Kettering, a cancer specialty hospital in New York, US. 4. The surgery occurred on October 30, 2020.",
"fact check": "1. Mukesh Ambani is the richest man in Asia. Not enough information given as the evidence does not mention anything about Ambani's wealth. 2. Mukesh Ambani had surgery for pancreatic cancer. Not enough information given as the evidence mentions a Facebook post but shortly after that he was seen at a launch event. 3. The surgery took place at Sloan Kettering, a cancer specialty hospital in New York, US. Not enough information given as the evidence does not mention anything about a hospital location. 4. The surgery occurred on October 30, 2020. The evidence shows other appearances by Ambani shortly before and after October 30, 2020. This contradicts with the fact that the surgery occurred on October 30, 2020.",
"support": 0,
"contradict": 1,
"facts count": 4
}}

Figure 2: Prompt we used for the Ev^2R reference-less prompt scorer. Given as input the claim and retrieved evidence, the scorer decomposes the claim and evaluates the resulting facts against the evidence data.

Proxy-Reference Prompt

You will get as input a claim and evidence.

Decide if the evidence supports the claim, refutes it, or doesn't give enough information. Explain the reasoning step-by-step before giving the answer. Only use the provided information and no additional sources or background knowledge.

Generate the output in form of a json as shown in the example below.

Examples:

Claim: South Africans that drink are amongst the top drinkers in the world.

Evidence: What is the global average alcohol consumption in litres of pure alcohol per day? The global averages as of 2016 is 15.1 litres per day. What is the daily average of pure alcohol consumption per day in South africa? 29.9 litres. Where does South Africa rank as a nation in terms of Daily pure Alcohol consumption? 6th out of 189 countries.

Output: {{
“explanation”: "The claim stays amongst the top drinkers not the top first, so since they are 6th, this could be plausible. The answer is support.",
“label”: 'supported'
}}

Claim: All government schools in India are being privatised.

Evidence: What did India's Union Education Minister say about the privatisation of governments schools? New Delhi: There is no plan to privatise primary education, the Centre told the Parliament today. This statement was given by Minister of Human Resource Development,
Ramesh Pokhriyal Nishank in the Lok Sabha today in response to Kaushalendra Kumar question on whether it is fact that NITI Aayog has suggested that Primary Education may be given to the private sector to reduce the burden of salary to teachers and other infrastructure.

Output: {{
“explanation”: "There is no plan by the Indian government to privatize primary education as said by the Minister of Human Resource Development. The claim is clearly refuted and therefore the answer is refute.",
“label”: 'refuted'
}}

Figure 3: Prompt we used for the Ev^2R proxy-reference prompt. We instruct the backbone LLM to reason over the retrieved evidence to predict a verdict label is assessed against the gold verdict label we use as proxy reference.

Reference-Based Prompt

You will get as input a claim, a reference evidence and a predicted evidence.
Please verify the correctness of the predicted evidence by comparing it to the reference evidence, following these steps:
1. Break down the PREDICTED evidence in independent facts. Each fact should be a separate sentence.
3. Evaluate each fact individually: is the fact supported by the REFERENCE evidence? Do not use additional sources or background knowledge.
4. Next, break down the REFERENCE evidence in independent facts. Each fact should be a separate sentence.
5. Evaluate each fact individually: is the fact supported by the PREDICTED evidence? Do not use additional sources or background knowledge.
5. Finally summarise (1.) how many predicted facts are supported by the reference evidence, (2.) how many reference facts are supported by the predicted evidence.

Generate the output in form of a json as shown in the examples below.

Examples:

Claim: Mukesh Ambani, richest man in Asia had surgery for pancreatic cancer at Sloan Kettering, New York, US cancer speciality hospital on October 30, 2020.

Reference evidence: When was the photograph taken of Mukesh Ambani, the richest man in Asia, on the Facebook post claiming he had been diagnosed with pancreatic cancer and had undergone surgery? The photograph was taken on September 5, 2020. When was a video filmed of Mukesh Ambani at the virtual launch of NK Singh's book Portrait of Power? The video was filmed on October 19, 2020. What date was the Facebook post which confirmed Mukesh Ambani had lost 30 kgs, been diagnosed with pancreatic cancer and had had liver transplant surgery? The Facebook post was dated November 2, 2020. Where was Mukesh's photo of him supposedly recieving surgery actually taken? It was taken by Manushree Vijayvergiya who shared her experience of meeting Mukesh and Isha Ambani in a cafe in Liechtenstein.

Predicted evidence: Who was the man who allegedly threatened Mukhesh Ambani: Question answer: What was his name and where was he from? New Delhi: The man had allegedly called on the HN Reliance Foundation Hospital and issued threats to Mukesh Ambani in filthy language. What are the predictions by Chiragh Darwalla for Ambanis? Astrology Predictions for Mukesh Ambani by Astrologer Chirag Daruwalla. Mukesh Ambani is an Indian industrialist and the chairman and managing director of Reliance Industries. What happened to Mukhesh Ambanis net worth? Nov 2, 2020 Mumbai: Mukesh Ambani, Asia's richest man, lost as much as $7 billion from his networth as Reliance Industries Ltd.'s shares tumbled to the lowest price in more than three months following a.

Output: {{
"facts in predicted evidence": "1. A man allegedly called on the HN Reliance Foundation Hospital and issued threats to Mukesh Ambani. 2. Astrologer Chirag Daruwalla issues predictions for Mukesh Ambani. 3. Mukesh Ambani is an Indian industrialist. 4. Mukesh Ambani is the chairman and managing director of Reliance Industries. 5. Mukesh Ambani is Asia's richest man. 6. Mukesh Ambani lost $7 billion from his networth as Reliance Industries Ltd.'s shares tumbled to the lowest price in more than three months.",
"fact check predicted evidence": "1. A man allegedly called on the HN Reliance Foundation Hospital and issued threats to Mukesh Ambani. The reference evidence does not mention anything about a man calling and threatening Mukesh Ambani. Not enough information. 2. Astrologer Chirag Daruwalla issues predictions for Mukesh Ambani. The reference evidence does not mention anything about an Astrologer giving predictions about Mukesh Ambani's future. Not enough information. 3. Mukesh Ambani is an Indian industrialist. The reference evidence does not mention that Mukesh Ambani is an Indian industrialist. Not enough information. 4. Mukesh Ambani is the chairman and managing director of Reliance Industries. The reference evidence does not mention that Mukesh Ambani is the managing director of Reliance Industries. Not enough information. 5. Mukesh Ambani is Asia's richest man. The fact 'Mukesh Ambani is Asia's richest man' is supported by the reference evidence. 6. Mukesh Ambani lost $7 billion from his networth as Reliance Industries Ltd.'s shares tumbled to the lowest price in more than three months. The reference evidence does not mention that Mukesh Ambani lost money or why he lost it. Not enough information.",
"facts count predicted evidence": 6,
"support predicted evidence": 1,
"facts in reference evidence": "1. Mukhesh Aambi is the richest man in Asia. 2. On September 5, 2020 a photograph of Mukesh Ambani was taken claiming he had been diagnosed with pancreatic cancer and had undergone surgery. 3. On October 19, 2020 a video of Mukesh Ambani was filmed at the virtual launch of NK Singh's book. 4. On November 2, 2020 a Facebook post was posted confirming that Mukesh Ambani had lost 30 kgs, been diagnosed with pancreatic cancer and had had liver transplant surgery. 5. A photo of Mukhesh Ambani supposedly recieving surgery actually taken in Liechtenstein.",
"fact check reference evidence": "1. Mukhesh Aambi is the richest man in Asia. The predicted evidence mentions that Mukhesh Ambani is Asia's richest man, this fact is hence supported. 2. On September 5, 2020 a photograph of Mukesh Ambani was taken claiming he had been diagnosed with pancreatic cancer and had undergone surgery. The predicted evidence does not mention anything about Mukhesh Ambani's cancer diagnosis or surgery. Not enough information. 3. On October 19, 2020 a video of Mukesh Ambani was filmed at the virtual launch of NK Singh's book. Predicted evidence does not mention Ambani attending any book launch. Not enough information. 4. On November 2, 2020 a Facebook post was posted confirming that Mukesh Ambani had lost 30 kgs, been diagnosed with pancreatic cancer and had had liver transplant surgery. The predicted evidence does not mention any of this. Not enough information. 5. A photo of Mukhesh Ambani supposedly recieving surgery was actually taken in Liechtenstein. The predicted evdience does not mention anything about a survey or Ambani being in Liechtenstein. Not enough information.",
"facts count reference evidence": 5,
"support reference evidence": 1
}}

Figure 4: Prompt we used for the Ev^{2}R reference-based prompt scorer. Both the reference evidence and the retrieved evidence are decomposed into atomic facts before assessing them against each other.

E Human Evaluation

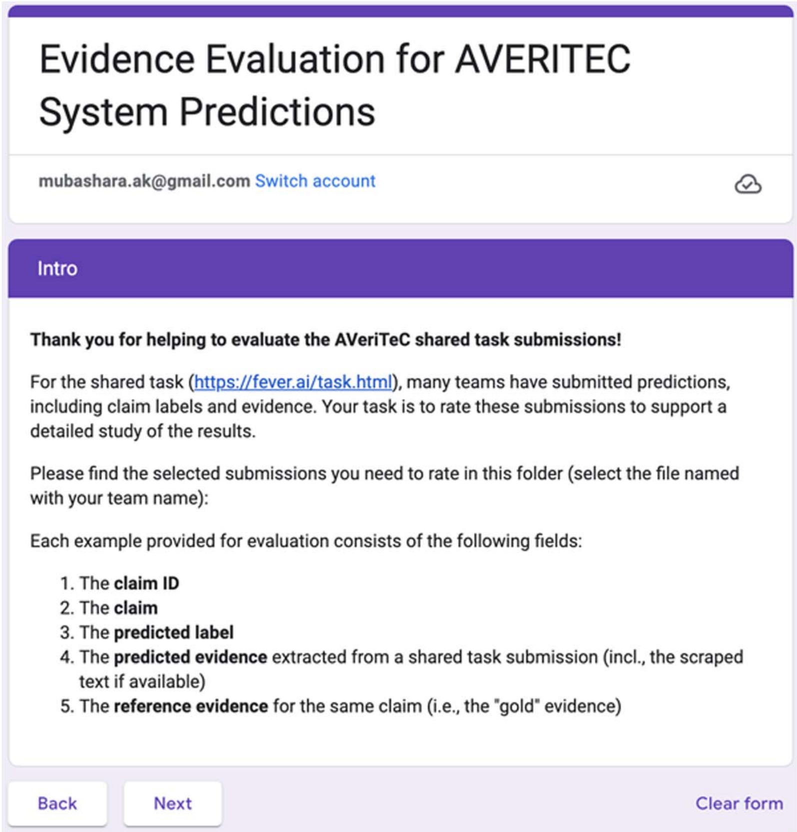
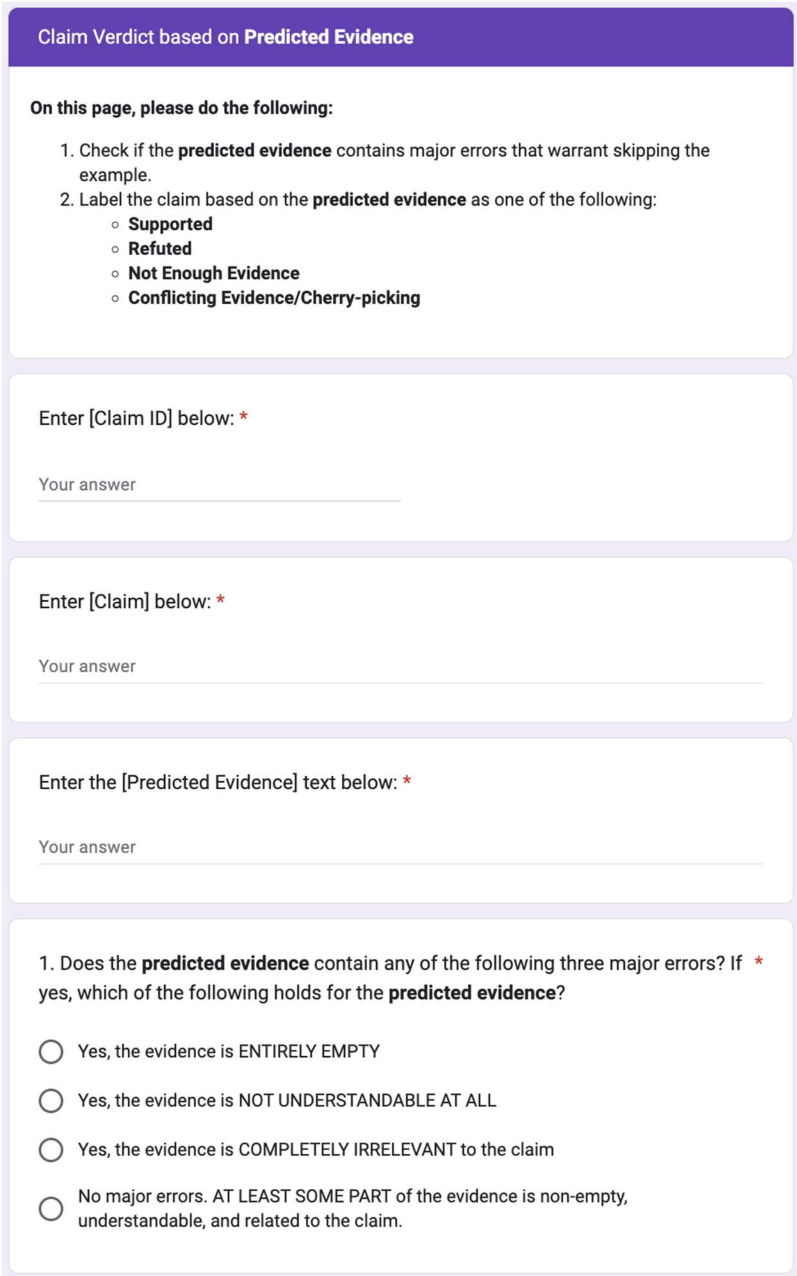


Figure 5: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.


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

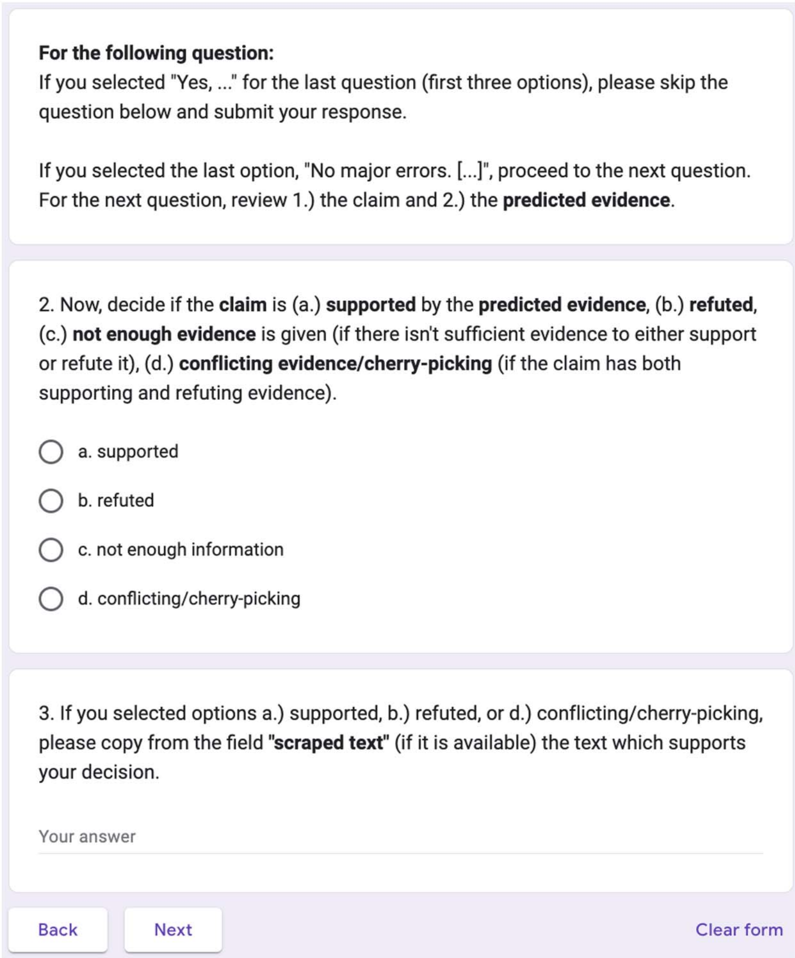
Enter the [Predicted Evidence] text below: *

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 6: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.


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

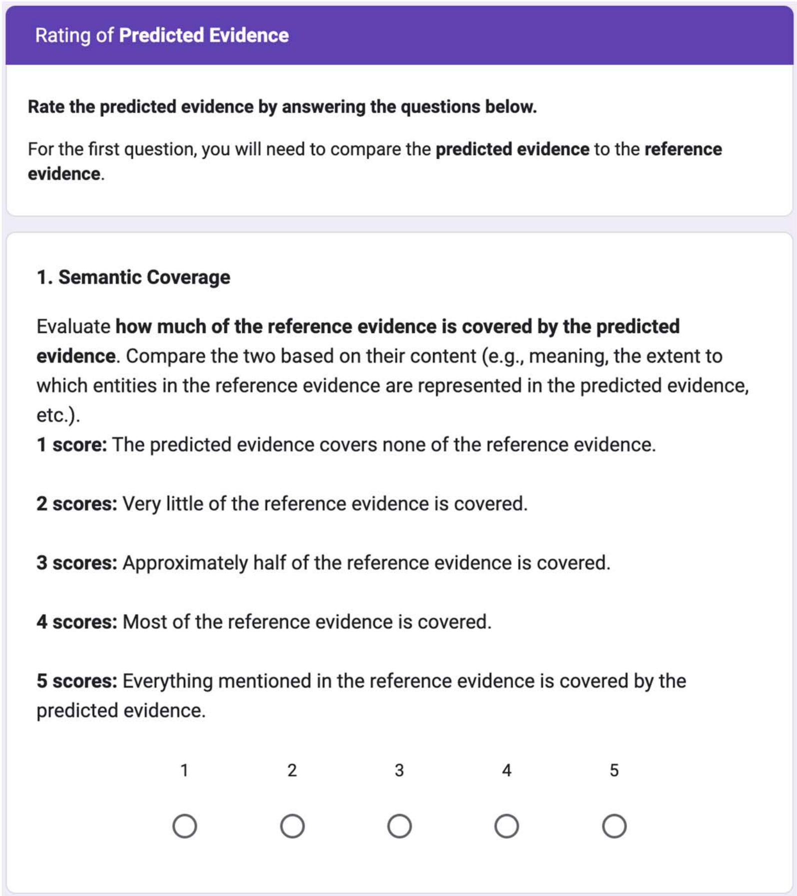
d. conflicting/cherry-picking

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 7: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.



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.

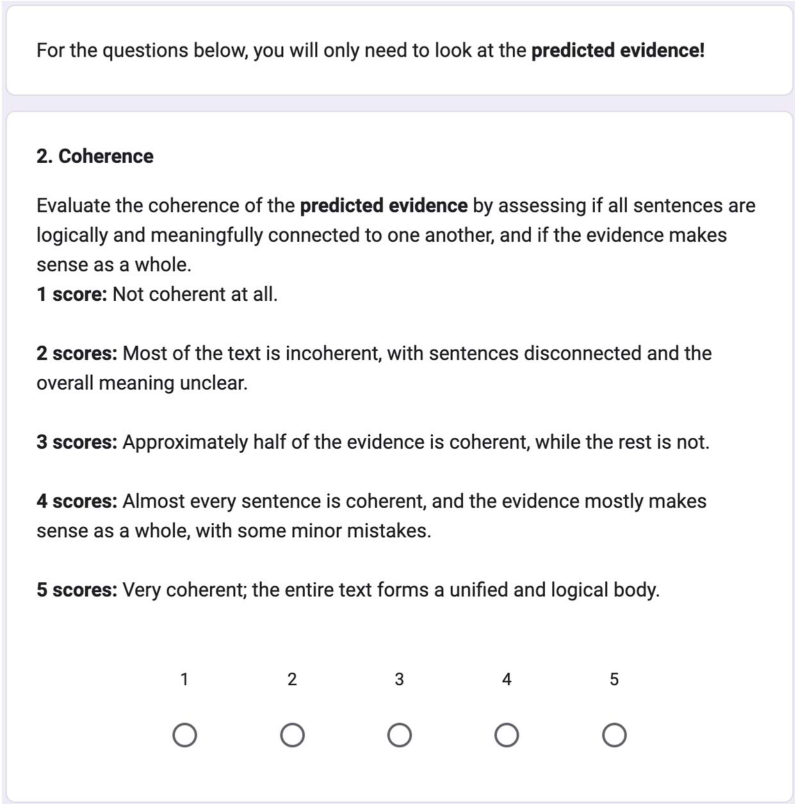
3 scores: Approximately half of the reference evidence is covered.

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 8: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.



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.

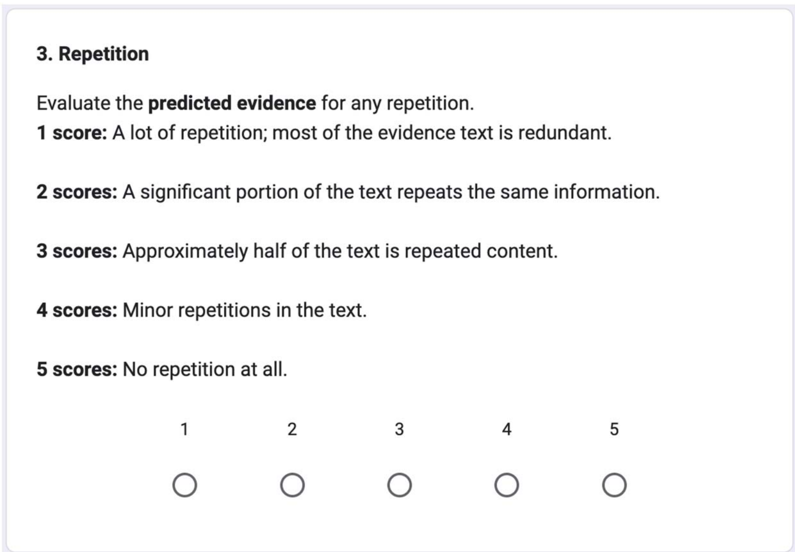
3 scores: Approximately half of the evidence is coherent, while the rest is not.

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 9: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.



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.

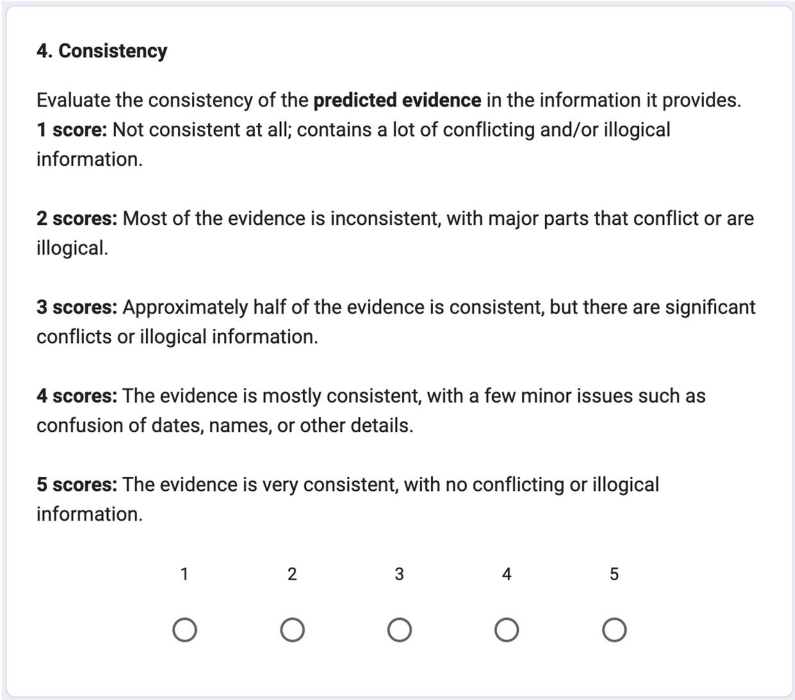
3 scores: Approximately half of the text is repeated content.

4 scores: Minor repetitions in the text.

5 scores: No repetition at all.

1 2 3 4 5

Figure 10: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.



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.

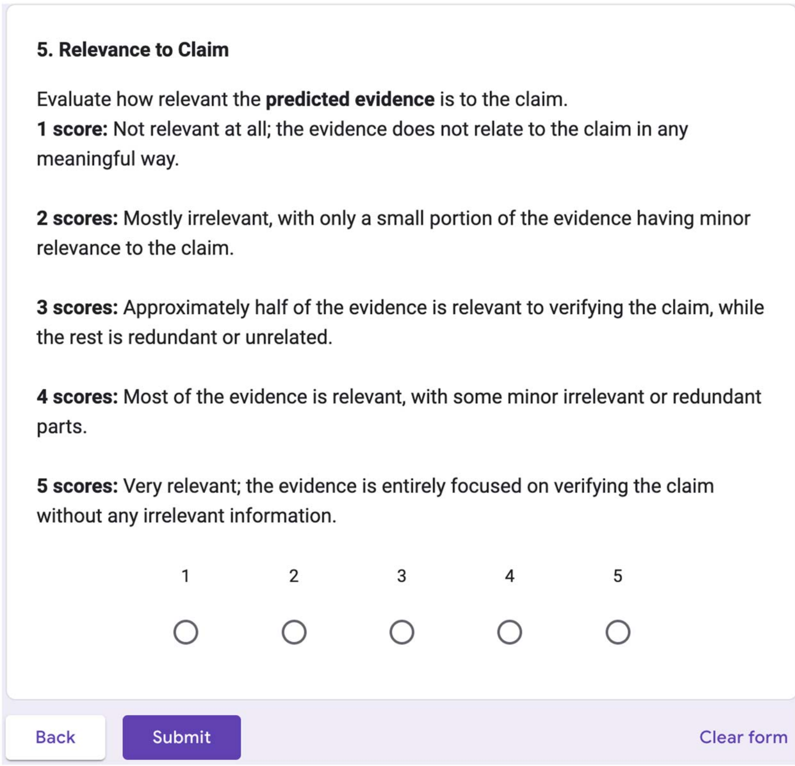
3 scores: Approximately half of the evidence is consistent, but there are significant conflicts or illogical information.

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

Figure 11: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.



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 12: Platform for human evaluation of retrieved evidence from systems submitted to the AveriTeC shared task.