SemQA: Evaluating Evidence with Question Embeddings and Answer Entailment for Fact Verification

Kjetil K. Indrehus
University of Oslo
kjetiki@ifi.uio.no

Caroline K. Vannebo
University of Oslo
carolikv@ifi.uio.no

Roxana Pop
University of Oslo
roxanap@ifi.uio.no

Abstract

Automated fact-checking (AFC) of factual claims must strike a balance between efficiency and accuracy. Although sophisticated frameworks such as Ev^{2}R offer strong semantic grounding, they often carry a heavy computational burden; on the contrary, simpler overlap- or one-to-one matching metrics are far less demanding, but frequently diverge from human judgments. In this paper, we introduce **SemQA**, a lightweight and accurate evidence scoring metric that combines transformer-based question scoring with bidirectional NLI entailment in answers. SemQA is then evaluated through correlation analysis with existing metrics, examination of representative examples, and human evaluations.

1 Introduction

Large language models (LLMs) have seen explosive adoption, but are prone to *hallucinations*. Xu et al. argue that this is a core limitation for LLMs. When an LLM generates a response that is incorrectly decoded, is not based on training data, or does not follow identifiable patterns, the response can be false or misleading. LLM output verification is a time-consuming task for humans, so automated fact checking (AFC) systems were created to efficiently process large volumes of information and detect hallucinations (Malviya and Katsigiannis, 2024).

The shared task FEVER (Fact Extraction and VERification)[1] has driven progress by providing a standardized framework and datasets for AFC systems to retrieve evidence and predict veracity labels. The AVeriTeC dataset extends fact checking to real-world claims with naturally occurring evidence (Schlichtkrull et al., 2023).

Traditional AFC evaluation methods often evaluate evidence solely based on predicted verdicts or by comparison of evidence retrieved with closed knowledge sources. Ev^{2}R (Akhtar et al., 2024) was introduced as an evaluation framework for AFC to counteract these limitations. In fact, Ev^{2}R outperforms many traditional evaluation approaches. However, being an LLM-driven framework, Ev^{2}R can be computationally intensive. Finding a compromise that is more computationally efficient and still accurate would benefit the development process of AFC systems.

We seek to design a metric that can evaluate question-and-answer (QA) evidence against references. Building on insights from Ev^{2}R, Hungarian METEOR (Kuhn, 1955), and soft weighting of question similarities, we propose **SemQA**. A **Sem**antic **Q**uestion and **A**nswer metric. Our work makes the following contributions;

- The design of SemQA, which combines question embeddings with natural language inference (NLI) answer entailment into a single tunable metric that is up to 5x faster than Ev^{2}R with correlation with human judgments.
- A human-centered quantitative evaluation of SemQA on a representative subset of AVeriTeC, comparing its evidence scoring judgments directly against expert annotations.

2 Related work

FEVER 2025 implements two main metrics for evidence evaluation: Ev^{2}R (Section 2.1) and Hungarian METEOR (Section 2.2). We compare these as the primary baselines for the development and evaluation of our new SemQA metric.

2.1 Ev^{2}R

Ev^{2}R has three different variations for evaluation (Akhtar et al., 2024); reference-based, proxy-based, and reference-less scorer. These scorers are evaluated on the basis of how well their predictions

[1] FEVER Workshop homepage: https://fever.ai/index.html

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 184–200

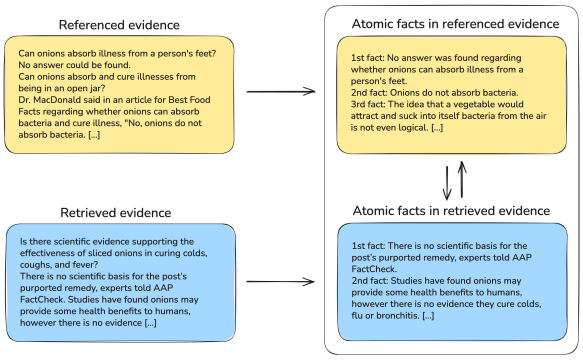


Figure 1: Example to visualize reference-based evaluation. The evidence is decomposed into atomic facts before evaluation. Illustration based on work of original author (Akhtar et al., 2024)

.

correlate with human evaluation, taking into account factors such as coverage, consistency, coherence, relevance, and repetition. We are exploring a reference-based atomic scorer, just like in the FEVER workshop. We deduce this on the basis of their implementation. The reference-based atomic scorer decomposes the retrieved evidence $\hat{E}$ and the referenced evidence E into atomic facts $A_{\hat{E}}$ and A_E. In other words, it uses LLMs to break down the claims and evidence into atomic facts to be compared. Figure 1 illustrates an example of this process.

The reference-based atomic scorer uses precision and recall scores. Precision refers to measuring the accuracy of the retrieved evidence, while recall is used to assess the completeness of the retrieved evidence $\hat{E}$ based on the gold standard. Akhtar et al. specifies the precision score s_{prec} as the ratio of facts supported by the referenced evidence:

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

The scorer iterates over each fact ($a_{\hat{E}} \in A_{\hat{E}}$), and if a fact $a_{\hat{E}}$ is supported by the referenced evidence E, then the indicator function ($\mathbf{I}[a_{\hat{E}}$ supported by $E]$) returns 1. In the opposite case, 0 will be returned. The recall score s_{recall} measures how much the retrieved evidence $\hat{E}$ covers the content of the referenced evidence E. Here, it evaluates whether each atomic fact of the referenced evidence ($a_E \in A_E$) is supported by the retrieved evidence $\hat{E}$ or not:

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

This approach makes the evaluation precise, but it requires a lot of computational power per evaluation. In practice, this is a significant limitation as the cost will scale with the size of the model and the number of claims. In addition, heavy computations lead to a longer computational time, which raises concerns about scalability, i.e. the performance of evaluation as the evidence corpus size grows. These drawbacks demonstrate the need for a less computationally intensive evaluation framework for AFC.

The evaluations in the paper of Ev^{2}R (Akhtar et al., 2024) suggest that the reference-based atomic scorer correlates better with human evaluations than traditional metrics. Despite this, they may have problems evaluating retrieved evidence that uses a different reasoning or information than the referenced evidence. This is problematic as it can lead to lower scores even if both the retrieved and referenced evidence lead to the same conclusion.

2.2 Hungarian METEOR

Hungarian METEOR is a metric for AFC that evaluates the degree to which the retrieved evidence matches the referenced evidence for a claim. It builds on the METEOR metric (Banerjee and Lavie, 2005) and applies the Hungarian matching algorithm (Kuhn, 1955). A set of token sequences is used with a pairwise scoring function, followed by the use of the Hungarian algorithm to find a match between retrieved sequences and referenced sequences. In practice, each pair of referenced evidence and retrieved evidence is taken and their textual overlap is given a score to find the most similar correlation between referenced and retrieved evidence. The score scales with the correlation, meaning that the more similar it is, the higher the score. Schlichtkrull et al. calculate the total score using $f(\cdot)$ as a pairwise scoring function. $X(\cdot)$ is a binary assignment function, where a match gives 1, and no match gives 0. The result u, is the maximum similarity score under the one-to-one matching restraint between the referenced (Y) and retrieved evidence ($\hat{Y}$):

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

Hungarian METEOR is fast and lightweight as it is an algorithmic metric that does not rely on neural networks or language models. This leads to a shorter computation time than the Ev^{2}R metric. However, this also means a limited semantic understanding. Akhtar et al. states that the use of token matching metrics, such as METEOR, is sensitive

to surface forms and does not consider alternative evidence paths.

3 Methodology

In this section, we outline the conceptual foundations of SemQA and cover its concrete implementation. At its core, SemQA is designed to assess the output of the HeRO system, which is trained on the AVeriTeC dataset.

There are various approaches to evaluate AFC systems. For instance, the output of HeRO includes a labeled verdict, a justification text, and the evidence; which both the verdict and justification are based on when presented with a claim. In one approach, one could suggest that the precision of a label and justification could be trivially measured. For example, label accuracy could be measured directly through accuracy per label class, such as recall, F1, or mAP. Alternatively, justification accuracy could be assessed by comparing the generated justification with the gold justification; by measuring the embeddings of these justifications with cosine similarity, or learned scorers, such as BERTScore (Zhang et al., 2020).

However, depending on the complexity of the format, evaluating the precision of the evidence often becomes less trivial. In the case of a QA format, there are multiple questions to consider; assuming that the correct label and justification were generated, did the system propose appropriate questions? Will the generated answers lead to the same justification as before? Could any generated answers contradict the gold justification?

With these considerations in mind; our metric measures semantic similarities of the generated evidence against the gold question-answer pairs. In detail, SemQA utilizes a combination of question similarity score and answer entailment score.

3.1 Formulation

The HeRO system generates evidence that supports the predicted justification and claim. The generated evidence $\hat{E}$, is a set of question-answer pairs, $\hat{P} = \{\hat{Q}, \hat{A}\}$, i.e., $\hat{E} = \{\hat{P}_0, \ldots, \hat{P}_n\}$. For reference-based evaluation, the gold QA pairs $P = \{Q, A\}$, labels L, and justifications J are provided, allowing us to directly evaluate performance against the generated output. The gold QA pairs can thus be evaluated directly against the predicted. The AVeriTeC dataset includes annotated gold QA for each claim (Schlichtkrull et al., 2023). However, the number of annotated questions is limited to a finite set; HeRO returns a set of generated questions that could be more than the number of annotated questions. This is taken into account with SemQA.

3.2 Question Score

Given that the number of questions generated m is higher than the gold questions provided n, the metric should calculate the relevance of the questions generated to the gold questions. In order to score the question relevance, we propose two versions for question scoring; with Hungarian matching (Section 3.2.1) and softmax (Section 3.2.2).

3.2.1 Variation: Hungarian Matching

Instead of Hungarian METEOR matching (Section 2.2), we utilize a sentence transformer to encode the question sentences into an encoded embedding $e(Q), e(\hat{Q})$. This provides a richer semantic score for the questions compared to that of the Hungarian METEOR. Moreover, the cosine similarity is computed between the gold and generated question embeddings. This is followed by building a similarity matrix based on each similarity score, $S_{i,j}$. Finally, the cost is calculated, $C_{i,j}$, turning similarity into a cost such that lower similarity results in higher cost:

$$S_{i,j} = \cos\big(e(Q_i),\, e(\hat{Q}_j)\big),$$
$$C_{i,j} = 1 - S_{i,j}.$$

Using the Hungarian matching algorithm (Kuhn, 1955), $\mathrm{HM}(\cdot)$, we can find an assignment of gold questions to generated questions with the lowest total cost. This gives us N pairs of lowest cost:

$$\mathrm{HM}(C) \;=\; \big\{(i,j)\big\}_{i=1}^{N}.$$

Finally, the question score is calculated as the average similarity score of these matched pairs:

$$Q_{\text{score}} \;=\; \frac{1}{N} \sum_{(i,j)\,\in\,\mathrm{HM}(C)} S_{i,j}.$$

3.2.2 Variation: Softmax

As an alternative to hard one-to-one matching, we can aggregate all pairwise question similarities via soft weighting. After encoding questions with the same transformer $e(\cdot)$, we form the cosine similarity matrix:

$$S_{i,j} \;=\; \cos\big(e(Q_i),\, e(\hat{Q}_j)\big)$$

We then normalize each row of S into a probability distribution:

$$P_{i,j} = \frac{\exp\big(S_{i,j}\big)}{\sum_{k=1}^{M} \exp\big(S_{i,k}\big)}.$$

Intuitively, $P_{i,j}$ measures how strongly the generated questions $\hat{Q}_j$ overlap with the gold questions Q_i. Let $\mathcal{F}$ be all probabilities over the set threshold t:

$$\mathcal{F} = \{(i,j) \mid P_{i,j} > t\}$$

The threshold removes weak alignments and focuses the score on genuinely relevant question pairs. This can also leave only a small number of surviving alignments. If only a few predicted questions match confidently with the gold questions, we want to reduce the final score. To improve this, we introduce the normalization constant k:

$$k \;=\; \min\Big(1,\; \frac{|\mathcal{F}|}{\min(N,M)}\Big),$$

k is calculated by taking the number of matches after thresholding, $|\mathcal{F}|$, over the maximum number of one-to-one matches, $\min(N,M)$. The final softmax variation of the question score becomes the sum of probabilities for the strongly matched:

$$Q_{\text{score}} \;=\; k \;\times\; \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} P_{i,j}.$$

3.3 Answer Score

When evaluating the quality of the generated answers, we only consider those tied to confidently matched questions, $\mathcal{F}$. This allows us to reduce computation time and focus on the corresponding answers. Let $\{(i,j)\}$ be the set of gold and predicted pairs of questions, returned by our question matching step. For each such pair (i,j), we extract the gold answer A_i and the generated answer $\hat{A}_j$, and run a bidirectional NLI, i.e., entailment in both directions. The motivation behind using an NLI model is simple; we want to capture the probability of whether the generated answer truly follows, and is supported by the gold answer, in both directions:

$$\begin{aligned} p_{\text{fwd}}(i,j) \;&=\; \text{Entail}\big(A_i \rightarrow \hat{A}_j\big) \\ p_{\text{bwd}}(i,j) \;&=\; \text{Entail}\big(\hat{A}_j \rightarrow A_i\big) \end{aligned}$$

$\text{Entail}(\cdot)$ is the probability of the "entailment" class after discarding the "neutral" dimension of NLI logits. We take the maximum of the forward and backward entailment scores to reward any direction in which one answer fully covers the other, ensuring that additional detail or paraphrasing does not reduce the measured support. Let $b_{i,j}$ be the strongest entailment for the given match i,j:

$$b_{i,j} \;=\; \max\big(p_{\text{fwd}}(i,j),\, p_{\text{bwd}}(i,j)\big).$$

Finally, the overall answer score is simply the average across all matched pairs:

$$A_{\text{score}} = \frac{1}{|\{(i,j)\}|} \sum_{(i,j)} b_{i,j}.$$

3.4 Weighted combination

Our metric balances question similarity and answer entailment with the hyperparameter α. By default, $\alpha = 0.5$. Setting a high α leads to more focus on question recall than answer entailment. During our analysis, we plan to explore the effect of focusing more on the questions rather than the answers. The metric is calculated as follows;

$$SemQA = \alpha \cdot Q_{\text{score}} \;+\; (1-\alpha) \cdot A_{\text{score}}$$

3.5 Implementation Details

For the sentence transformer, we utilize the pretrained; `All Mpnet Base V2`[2]. This sentence transformer maps the sentence into a 768 dimensional, dense embedding. We selected the model due to its balanced trade-off between computational efficiency and semantic richness, making it well-suited for our evaluation metric. Then, for our NLI model, we utilize `BART` (Lewis et al., 2019).

4 Analysis and Results

We assess our metric through three forms of analyses. First, in Section 4.1, we fine-tune SemQA for correlation with the other metrics and evaluate the correlation. This is followed by computational efficiency in Section 4.2. Next, in Section 4.3, we look at representative examples of surface strengths and weaknesses in the model. Finally, in Section 4.4, we hold human evaluations for further examination of our metric.

4.1 Finetuning for Correlation

Our implementation has multiple tunable parameters and variations. This includes variations in question scores (Section 3.2), alpha α, and threshold. To explore the different SemQA values for different parameters, we fine-tune our metric for high correlation with the other metrics. First, leveraging the `HeRO` system (Yoon et al., 2024) with the instruction-tuned `Llama 3.3 70` model as a *"judge"* for evidence generation (Meta, 2024), we generate examples. Then, the generated output is utilized for both fine-tuning and evaluation.

[2]Link to sentence transformer: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

To fine-tune our metric, we explored different alpha and threshold values for the metrics. Subsequently, we were interested in investigating the effect of the implemented variations. To evaluate SemQA, we compared it with the other metrics by Covariance, Pearson's r and Kendall's τ. Covariance is used to evaluate how the two metrics co-vary around their means, while Person's r correlation coefficient r measures the strength of a linear relationship between two normally distributed variables (Benesty et al., 2009). Kendall's τ assesses the ordinal association between two rankings by counting concordant and discordant pairs, providing a nonparametric measure of monotonic relationship that is robust to non-Gaussian distributions (Kendall, 1938). In addition, we performed a grid search on the `train-200` dataset (Table 1). The results are presented in Table 2.

DATASET	EXTRACTED SUBSET	EXAMPLES
`Train`	✗	3068
`Train-200`	✓	200
`Dev`	✗	500

Table 1: Overview of the dataset sizes for the project.

Table 2 demonstrates how SemQA correlates with four *off-the-shelf* metrics under five representative hyperparameter settings. In each row, we report on covariance, Pearson's r, and Kendall's τ for SemQA compared to baseline metrics for the `AVeriTeC` dataset (Schlichtkrull et al., 2023). The findings show a consistent alignment of the SemQA Hungarian matching variant with pure question recall (peak $r \approx 0.79$ and $\tau \approx 0.61$), in addition to maintaining moderate correlation with the Hungarian QA recall score ($r \approx 0.43$, $\tau \approx 0.26$).

In contrast, the softmax variant shows weaker correlations (peak $r \approx 0.45$, $\tau \approx 0.28$). Even after thresholding, the softmax variant still spreads the probability mass across all remaining pairs instead of focusing on a single, strongest match. Further analysis is done with the Hungarian variation of the question score ($\alpha = 0.8$ and threshold $= 0.2$).

4.2 Computational Efficiency

Table 3 presents the computation times for the different metrics. The results demonstrate that SemQA requires substantially less computational time than Ev^{2}R. However, it is still more expensive than pure surface-based approaches, such as Hungarian METEOR. These findings meet our expectations and suggest that SemQA is a suitable and less computationally intensive alternative to Ev^{2}R.

METRIC	TIME PER EXAMPLE (S)
Hungarian METEOR (Q)	**0.0374**
Hungarian METEOR (Q+A)	0.0738
AVeriTeC end-to-end	0.0697
Ev^{2}R Q-only recall	7.0180
Ev^{2}R Q+A recall	7.3836
SemQA	1.4935

Table 3: Average computation time per claim for each metric. Computed by using `NVIDIA A100 40GB PCIe` on the training subset (200 examples). Lowest computation time highlighted in bold.

4.3 Manual Evaluation

To explore how well SemQA captures semantic similarities of the evidence, we sampled and investigated five edge cases illustrated in Tables 4-7; the full examples are located in Appendix A Tables 11-14. We compare SemQA scores with the baseline metrics; Hungarian METEOR, Ev^{2}R Q-only recall and Ev^{2}R QA recall. We are interested in whether or not SemQA appropriately penalized or rewarded the retrieved evidence against the referenced evidence based on its meaning.

SAMPLE 16 (TAXES)	
GOLD QA	
Q1:	Has tax revenue risen since taxes were lowered in 2017
A1:	No, Really, Tax Revenue Has Not Risen
Q2:	What is the value of the total tax revenue in the 2017/2018 fiscal year since the TCJA was signed into law
A2:	Total revenue over the time period in question has actually fallen by 1.6 percent in real (inflation-adjusted) terms
GENERATED QA	
Q2:	Did the 2017 tax cuts lead to an increase in Treasury revenues?
A2:	The most recent CBO projections estimate further decreases in corporate tax revenue. The TCJA also reduced income taxes for most Americans, which led to a decline in revenues relative to prior projections.
Q6:	What was the change in payroll taxes after the tax cuts in 2017
A6:	In fact, payroll taxes fell only slightly—1.7%—from pre-TCJA projected values. This provides baseline credibility that reinforces the declines in other revenues.
Hungarian METEOR: **0.359**	Ev^{2}R Q-only: **0.50**
Ev^{2}R Q+A: **1.0**	SemQA: **0.689**

Table 4: Sample 16 with claim: *"We actually saw revenues to the Treasury increase after we lowered taxes in 2017. Rest assured the Democrats"*. The example shows that the metrics are not closely aligned. Full example in Appendix A Table 11.

In Table 4 we observe a moderately strong se-

α	threshold	variation	Q-only (Hungarian)			QA (Hungarian)			Ev^{2}R Q-only			Ev^{2}R QA		
			Cov	r	τ	Cov	r	τ	Cov	r	τ	Cov	r	τ
0.7	0.8	hungarian	0.020	0.714	0.517	0.010	0.434	0.262	0.022	0.377	0.248	0.020	0.362	0.200
0.9	0.2	hungarian	0.022	0.793	0.606	0.010	0.433	0.289	0.025	0.442	0.310	0.017	0.308	0.198
0.8	0.8	softmax	0.002	0.251	0.179	0.001	0.139	0.067	0.001	0.065	0.030	0.008	0.414	0.374
0.8	0.3	softmax	0.021	0.449	0.284	0.010	0.261	0.100	0.030	0.294	0.190	0.019	0.204	0.131
0.3	0.8	softmax	0.007	0.251	0.179	0.004	0.139	0.067	0.004	0.065	0.030	0.028	0.414	0.374

Table 2: Covariance, Pearson's r, and Kendall's τ between the SemQA composite score and each baseline metric under selected hyperparameter settings.

mantic match between the QA pairs which SemQA scores as 0.68, while Hungarian METEOR underestimates (0.35). The difference in Ev^{2}R QA (1.0) and Q-only (0.5) recall makes sense, due to the partial semantic match of generated questions and the strong alignment of answers. Compared to baseline metrics, SemQA rewards meaningful evidence paths more appropriately. In addition, this example demonstrates SemQA's recognition of paraphrases and partial entailment despite low token overlap. When considering the textual content of the QA pairs, we evaluate this score as accurate. Another example of this can be found in the Appendix A Table 10.

SAMPLE 120 (HARRIS STATEMENT)

GOLD QA

Q2: Where was the claim first published
A2: The claim was first published on bustatroll.org
Q3: What kind of website is bustatroll.org
A3: It is a satirical website and identifies as a subsidiary of America's Last line of Defense' network of parody, satire, and tomfoolery

GENERATED QA

Q2: Was the statement attributed to Kamala Harris actually made by her?
A2: Facebook users expressed outrage over a statement that's falsely attributed to Sen. Kamala Harris. The quote — about seeking "vengeance" against supporters of President Donald Trump — originated on a website that calls its work satire.
Q10: Is the quote attributed to Kamala Harris a genuine statement or a fabrication?
A10: Reuters, "Fact check: Kamala Harris quote on 'vengeance of a nation' fabricated by satire website," June 25, 2020

Hungarian METEOR: **0.166** | Ev^{2}R Q-only: **0.0**
Ev^{2}R Q+A: **0.0** | SemQA: **0.253**

Table 5: Sample 120 with claim: *"Kamala Harris said "And once Trump's gone and we have regained our rightful place in the White House, look out if you supported him and endorsed his actions, because we'll be coming for you next".* All metrics show low scores. Full example in Appendix A Table 12.

Table 5 is an interesting case as all baseline metrics give poor scores (Ev^{2}R 0.0, Hungarian METEOR 0.16). The slightly higher SemQA score (0.25) is likely due to minimal overlap of the keyword *"satire"*. Demonstrating once again its ability to find semantic correlations.

In the case of highly semantic similarities, the generated answers in Table 6 mirror the gold answer sentence "I don't support defunding police". SemQA and both Ev^{2}R metrics gave an impressive score of 1.0, demonstrating that any rephrasing of the evidence was correctly captured and interpreted. This type of successful entailment is not accessible to Hungarian METEOR (0.37), which lags behind the other metrics in semantics.

SAMPLE 10 (BIDEN STATEMENT)

GOLD QA

Q1: Does Joe Biden support defunding the police?
A1: NO he said: "Federal dollars should not go to departments that violate people's rights or turn to violence as a first resort, but I doń support defunding police."

GENERATED QA

Q5: Does Joe Biden support defunding the police?
A5: "I do not support defunding police," Biden wrote in an op-ed for USA Today. "The better answer is to give police departments the resources they need to implement meaningful reforms, and to condition other federal dollars on completing those reforms.
Q7: Does Joe Biden support defunding the police?
A7: Presumptive Democratic nominee Joe Biden definitively declared "I do not support defunding police," in an op-ed Wednesday, as protesters around the country increase their calls for overhauling the criminal justice system and President Trump attempts to tie Biden to the "Defund the police" movement.

Hungarian METEOR: **0.375** | Ev^{2}R Q-only: **1.0**
Ev^{2}R Q+A: **1.0** | SemQA: **1.0**

Table 6: Sample 10 with claim: *"Biden has pledged to defund the police".* All metrics show high score, and indicate an agreement. Full example in Appendix A Table 13.

Finally, Table 7 shows the gold QA stating that the Sputnik vaccine has not been thoroughly tested. The generated evidence mentions the status of the vaccine in terms of trials, registration, and safety concerns, but is not as explicit as the gold QA. One of the generated answers even states *"the vaccine developed by the Gamaleya Institute in Moscow is*

safe" as part of a longer sentence, where the writing style can be misleading when not reading the full context. SemQA (0.47) sits between Hungarian METEOR (0.29) and the Ev^{2}R metrics (0.6/0.5), reflecting its nuanced partial mismatch.

SAMPLE 121 (RUSSIAN VACCINE)

	GOLD QA
Q1:	Who has developed this vaccine?
A1:	'Sputnik V' has been developed by Moscow-based Gamaleya Research Institute of Epidemiology and Microbiology.
Q4:	Has it been thoroughly tested?
Q5:	Has it been tested for safety?
A4-A5:	No
	GENERATED QA
Q5:	Has the COVID-19 vaccine developed by Russia been proven to be safe and effective?
A5:	Scientists and public health officials are skeptical about Russian President Vladimir Putin's claim that the country's potential vaccine for the coronavirus "works quite effectively," saying Tuesday that the vaccine still needs critical testing to determine whether it's safe and effective.
Q8:	Has the COVID-19 vaccine developed by the Gamaleya Institute in Moscow been thoroughly tested?
A8:	Despite having only been in clinical trials for less than two months, the vaccine developed by the Gamaleya Institute in Moscow is safe, Putin said at a televised cabinet meeting, noting that it has already been given to one of his daughters, according to Reuters and The Washington Post.

Hungarian METEOR: **0.294**	Ev^{2}R Q-only: **0.6**
Ev^{2}R Q+A: **0.5**	SemQA: **0.476**

Table 7: Sample 121 with claim: *"Russia has successfully developed a vaccine for Covid-19 and it has passed all checks."*. SemQA aligns closely with Ev^{2}r, while Hungarian METEOR is much lower. Full example in Appendix A Table 14.

In summary, the sample in Table 4 shows that SemQA is able to identify evidence that is semantically correct but lexically divergent. This aligns with our expectations. SemQA measures deep semantic similarity instead of simple n-gram overlap as in Hungarian METEOR. SemQA successfully assigns low scores when the generated evidence simply mentions relevant terms without substantially matching the gold evidence, as we saw in Table 5. In borderline cases where evidence is partially similar but missing critical nuances, SemQA produced midrange scores that reflect partial support, illustrated in Table 7. This precision of false or misleading claims is highly relevant when evaluating AFC systems. Finally, Table 6 demonstrates SemQA's ability to identify the gold fact in different words that are semantically equivalent; reaching a full score of 1.0. The manual evaluation suggests that SemQA's use of sentence embeddings and entailment scoring is capturing semantic similarities by rewarding correct paraphrasing and penalizing insufficient evidence.

4.4 Human Evaluation

AFC systems rely on quantitative metrics to evaluate performance, but these metrics do not consider nuances and might give unrepresentative scores. Human feedback can point out shortcomings and recognize the difference between harmless and critical mistakes made. Human evaluations allows for a more qualitative analysis. Due to this, we wanted to investigate the accuracy of SemQA from a human point of view.

As SemQA calculates a score from 0-1 based on how semantically aligned the retrieved evidence is with the referenced evidence, it made sense to collect ordinal data to analyze this accuracy. Each evaluation set consisted of 10 examples of referenced evidence, retrieved evidence, and the SemQA score. For each example, the participants evaluated how accurate the SemQA score was on a scale of 1-7, where 1 = *"score should be much lower"*, 4 = *"score is accurate"*, and 7 = *"score should be much higher"*. This is illustrated by the instructions in Figure 3, and the example in Figure 4 in Appendix B.

In total, we had 22 participants with a background in computer science, informatics, or information technology of varying degrees. Most of the participants (15) are professionals working in the industry as developers, architects, or engineers. Including both in-house and consultant roles. The rest of the participants (7) are postgraduate students. All participants reside in the Oslo region, the majority of them being male (14), with only 8 female participants. We assigned the same evaluation set to a pair of participants, i.e. two annotators per evaluation set. Having 22 participants, this led to 11 different evaluation sets and a total of 110 examples.

Figure 2 illustrates a histogram of the frequency of each evaluation score, regardless of the example being evaluated. It illustrates the distribution of scores in human evaluation.

In the histogram, the majority of human evaluation scores are below 4.0, making up 48.64%. This means that according to the annotators, the SemQA score was too high. In other words, SemQA may sometimes over-reward evidence according to human evaluations. Only a minority of the evaluations marked the SemQA score as being too low; with

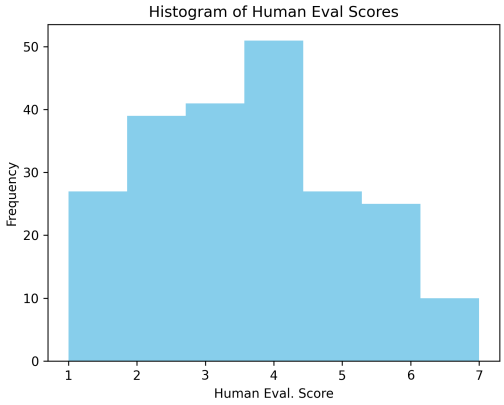


Figure 2: Histogram of frequency of human evaluation scores.

28.18% of human evaluation scores above 4.0. This suggests that SemQA over-rewards the evidence, not aligning itself with human evaluations.

In the histogram, the most frequent score is 4.0, making up 23.18%. This score marks the SemQA score as accurate. Human evaluation scores of 3.0-5.0 make up 54.09% of the distribution, indicating that the SemQA score was accurate or close to accurate.

To investigate this further, we made comparisons of individual evaluations. In Table 8 we see a SemQA score of 0.69 and human evaluation scores of 5 and 6. This means that the annotators agreed that the SemQA score should have been slightly to moderately higher. It is the opposite case for the SemQA score of 0.94; here the annotators gave a score of 1 and 2, both evaluating the SemQA score as too high. Then, the annotators disagree; one classified the SemQA score of 0.78 as too low (6), while the other annotator evaluated the SemQA score as accurate (4). These findings led us to further examine the consensus between the annotators.

SemQA	Ann.1	Ann.2	Consensus
0.69	6	5	Agreement
0.94	1	2	Agreement
0.78	4	6	Disagreement

Table 8: Comparison of individual human evaluations. Visual representations of these findings are supplemented in Appendix B Figures 5-7.

We found that 21 of 110 evaluations were in total agreement, representing 19.09%. The number of agreements with a tolerance of 1 made up 57 of 110 evaluations, or 51.82%. 85 of 110 evaluations were in agreement with a tolerance of 2, making up 77.27%.[3] This distribution shows a trend of a majority of annotators in relative agreement with each other when evaluating the accuracy of SemQA scores.

From the results of human evaluations, we calculated the mean, standard deviation, and median of all human evaluations. Table 9 shows a mean of 3.57, which is very close to 4.0. With this result, we interpret that according to human evaluations, SemQA is relatively accurate and manages to capture semantic similarities. The standard deviation is small, which tells us that the annotators agreed that the SemQA score was accurate.

HUMAN EVALUATION SCORES		
Mean	std	Median
3.5773	1.6756	4

Table 9: Calculations of human evaluation scores. Calculated by collecting all 220 human evaluation scores.

5 Conclusion

Our proposed metric, SemQA, is a reference-based metric that evaluates based on question-answer pairs. We show that a weighted question-and-answer score can be used to evaluate the evidence. SemQA relies on a sentence transformer and NLI model; where it is still able to compute five times faster than Ev^2r, while aligning with the Ev^2r metric. Our human evaluation confirms that SemQA reflects evidence quality more faithfully than overlap-based baselines. We believe that SemQA provides a practical, efficient, and reliable metric for the development and evaluation of automated fact-checking systems.

Limitations

Our metric assumes that the given evidence is based on question-and-answer pairs, i.e referenced based. Factual justification does not need to be in question-and-answer form only. In a different context, it makes sense to extract information into a summary of evidence. For example, `JustiLM` generates multi-sentence justifications by retrieving and synthesizing evidence into fluent text, rather than question-answer pairs (Zeng and Gao, 2024).

Our metric is a reference-based evaluation framework of questions and answers. The dependence

[3]Tolerance in this context refers to scores above or below the referenced one. For example, a score of 4 with a tolerance of 1 refers to scores 3-5.

of references are tightly in the design of the metric. Without annotated gold question and answers, applying our metric would require to extract gold question and answers first.

Further Work

Further work might explore ways to generalize the metric to other formats beyond question-and-answer pairs. Such formats could be free-text or structured multi-sentence evidence. This would make the metric applicable to a larger number of fact-verification systems.

References

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev^{2}r: Evaluating evidence retrieval in automated fact-checking. *Preprint*, arXiv:2411.05375.

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.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Shrikant Malviya and Stamos Katsigiannis. 2024. SK_DU team: Cross-encoder based evidence retrieval and question generation with improved prompt for the AVeriTeC shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 99–107, Miami, Florida, USA. Association for Computational Linguistics.

Meta. 2024. Meta llama 3.3 70b. Release Date: December 6, 2024.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Preprint*, arXiv:2305.13117.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. Hero at averitec: The herd of open large language models for verifying real-world claims. *Preprint*, arXiv:2410.12377.

Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Preprint*, arXiv:2401.08026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Examples from Analysis

Example 89 (Mail-in Ballot)

Claim: In the US, envelopes used for mail-in ballots in general elections reveal the party affiliation of the voter allowing postal workers to toss votes.

Gold QA

Q1: Do ballot envelopes in general elections, show who the voter voted for?
A1: No

Q2: When are indications shown on the envelope?
A2: The office of the supervisor of elections for Palm Beach County, Florida — where the envelopes in the Facebook posts were sent — confirmed that in a statement to FactCheck.org. In a closed primary, "voters are sent ballots according to their precinct and their party affiliation. The ballots in that video are primary ballots." The alphanumeric codes on the envelopes shown in the video are used to direct the ballots for counting.

Generated QA

Q1: Can postal workers tamper with mail-in ballots in the US?
A1: (2/3) each other. Because of this, voters are sent ballots according to their party affiliation. It would be a federal crime for any post office employee to tamper with this process and we have received no such reports of anything like this happening.

Q2: Do envelopes used for mail-in ballots in general elections in the US include any information that could reveal the party affiliation of the voter?
A2: Return envelopes for all election ballots must include prepaid postage. Service and overseas voters must be provided with instructions and a privacy sheet for returning the ballot and signed declaration by fax or email.
Q3: What information is verified by the SVC when there is an SVC on an envelope ballot by mail?
A3: envelope ballots by mail. If there is a SVC, the SVC must verify the signatures before the EVBB

Q4: Are the envelopes used for mail-in ballots in the US sealed to prevent postal workers from seeing the party affiliation of the voter?
A4: that all ballot envelopes are sealed or that authorized personnel accompany the ballots to the Post Office or
Q5: Are mail-in ballots in general elections in the US sent in secrecy envelopes?
A5: secrecy envelopes, if used by the voter. (The EVBB has the discretion to open a ballot secrecy
Q6: Do mail-in ballots in the US guarantee the secrecy of a voter's ballot?
A6: A voter who returns a ballot electronically is deemed to have acknowledged that there may be situations where the secrecy of that voter's ballot cannot be guaranteed.
Q7: Do envelopes used for mail-in ballots in general elections in the US reveal the party affiliation of the voter?
A7: absentee ballot. However, applications and ballot envelopes of the temporarily or permanently disabled are not
Q8: How do the return envelopes for mail-in ballots in the US, as described in 39 U.S.C. 3406, affect the secrecy of the vote?
A8: Return envelopes for overseas and service voters must enable the ballot to be returned postage free if mailed through the United States postal service, United States armed forces postal service, or the postal service of a United States foreign embassy under 39 U.S.C. 3406.
Q9: How do states ensure the secrecy of mail-in ballots in the US?
A9: States Postal Service (USPS). The voter must be notified if the return envelope
Q10: Do envelopes used for mail-in ballots in general elections in the US reveal the party affiliation of the voter?
A10: (1) The county auditor shall send each voter a ballot, a security envelope in which to conceal the ballot after voting, a larger envelope in which to return the security envelope, a declaration that the voter must sign, and instructions on how to obtain information about the election, how to mark the ballot, and how to return the ballot to the county auditor.

Scores:

Hungarian METEOR	Ev^{2}R Q-only	Ev^{2}R Q+A	SemQA
0.373	0.50	0.667	0.548

Table 10: Example 89. Claim: *"In the US, envelopes used for mail-in ballots in general elections reveal the party affiliation of the voter allowing postal workers to toss votes".* This example shows that the metrics are closely aligned.

Example 16 (Taxes)

Claim: We actually saw revenues to the Treasury increase after we lowered taxes in 2017. Rest assured the Democrats do not want you to know that.

Gold QA

Q1: Has tax revenue risen since taxes were lowered in 2017
A1: No, Really, Tax Revenue Has Not Risen

Q2: What is the value of the total tax revenue in the 2017/2018 fiscal year since the Tax Cuts and Jobs Act (TCJA) was signed into law
A2: Total revenue over the time period in question has actually fallen by 1.6 percent in real (inflation-adjusted) terms

Generated QA

Q1: Did the tax cuts in 2017 lead to an increase in Treasury revenues in the following year?
A1: While some TCJA supporters observe that nominal revenues were higher in fiscal year 2018 (which began Oct. 1, 2017) than in FY2017, that comparison does not address the question of the TCJA's effects.

Q2: Did the 2017 tax cuts lead to an increase in Treasury revenues?
A2: The most recent CBO projections estimate further decreases in corporate tax revenue. The TCJA also reduced income taxes for most Americans, which led to a decline in revenues relative to prior projections.

Q3: According to the Congressional Budget Office, what were the projected corporate tax revenues for fiscal years 2018 and 2019 before the 2017 tax cuts?
A3: Several months before the TCJA was enacted, the Congressional Budget Office (CBO) projected that corporate tax revenues for fiscal years 2018 and 2019 would total $668 billion.

Q4: Did the actual amount of revenue collected by the US Treasury in FY2018 match the Congressional Budget Office's (CBO) projection made in January 2017?
A4: In fact, the actual amount of revenue collected in FY2018 was significantly lower than the Congressional Budget Office's (CBO) projection of FY2018 revenue made in January 2017—before the tax cuts were signed into law in December 2017.

Q5: Did the 2017 tax cut lead to an increase in tax revenue collected by the Treasury?
A5: The actual amount of tax revenue collected in FY2018 was significantly lower than the CBO's projection made in January 2017—before the tax cut was signed into law.

Q6: What was the change in payroll taxes after the tax cuts in 2017?
A6: In fact, payroll taxes fell only slightly—1.7%—from pre-TCJA projected values (Figure 2). This provides baseline credibility that reinforces the declines in other revenues.

Q7: How did the actual tax revenue in FY2018 compare to the predicted tax revenue in FY2018 if the Tax Cuts and Jobs Act (TCJA) had not been passed?
A7: The most appropriate test of the revenue impact of the TCJA is to compare actual revenues in FY2018 with predicted revenues in FY2018 assuming Congress had not passed the legislation.

Q8: According to the CBO, what was the projected change in corporate tax revenue after the 2017 tax cuts?
A8: In the forecast published soon after the TCJA was enacted, however, the CBO projected $519 billion in corporate tax revenue over those two years—a $149 billion decrease.

Q9: Did corporate tax revenue decrease after the 2017 tax bill?
A9: Since the TCJA was enacted, corporate tax revenue has been down from its projected level by about one-third, even as pretax corporate profits have continued to rise toward historic highs.

Q10: How did the 2017 tax cuts affect federal revenue according to the evidence?
A10: The individual income tax changes are scheduled to expire after December 31, 2025. If permanent, the income tax provisions would reduce federal revenue by $165 billion per year on a conventional basis, but when incorporating economic growth and feedback, on a dynamic basis, they would reduce federal revenue by $115 billion a year.

Scores:

Hungarian METEOR	Ev^2R Q-only	Ev^2R Q+A	SemQA
0.359	0.5	1	0.689

Table 11: Example 16. Claim: *"We actually saw revenues to the Treasury increase after we lowered taxes in 2017. Rest assured the Democrats".* The example shows that the metrics are not very well aligned with each other.

Example 120 (Harris statement)

Claim: Kamala Harris said "And once Trump's gone and we have regained our rightful place in the White House, look out if you supported him and endorsed his actions, because we'll be coming for you next".

Gold QA

Q1: When was the claim made
A1: 11/08/2020

Q2: Where was the claim first published
A2: The claim was first published on bustatroll.org

Q3: What kind of website is bustatroll.org
A3: It is a satirical website and identifies as a subsidiary of America's Last line of Defense' network of parody, satire, and tomfoolery

Generated QA

Q1: Did Kamala Harris make a statement about punishing Trump supporters after his impeachment?
A1: "Kamala Harris: 'After We Impeach, We Round Up The Trump Supporters,'" the story's headline reads. While Harris has supported beginning impeachment proceedings against Trump, the story presents an otherwise imagined tale about a Harris speech.

Q2: Was the statement attributed to Kamala Harris actually made by her?
A2: Facebook users expressed outrage over a statement that is falsely attributed to Sen. Kamala Harris. The quote — about seeking "vengeance" against supporters of President Donald Trump — originated on a website that calls its work satire.

Q3: What did Kamala Harris say would happen to Trump supporters after he leaves office?
A3: CLAIM: Kamala Harris, a U.S. senator and former Democratic presidential candidate from California, said on June 18 that once President Trump is no longer in office "and we have regained our rightful place in the White House," his supporters will feel the "vengeance of a nation."

Q4: What did Kamala Harris say would happen to Trump supporters after he leaves office?
A4: CLAIM: Kamala Harris, a U.S. senator and former Democratic presidential candidate from California, said on June 18 that once President Trump is no longer in office "and we have regained our rightful place in the White House," his supporters will feel the "vengeance of a nation."

Q5: Did Kamala Harris make a statement threatening to target people who supported Trump after he leaves office?
A5: A statement first attributed to Democratic Sen. Kamala Harris in a made-up story is now making the rounds on social media as if it's real.

Q6: Did Kamala Harris say "And once Trump's gone and we have regained our rightful place in the White House, look out if you supported him and endorsed his actions, because we'll be coming for you next"?
A6: Though Harris has been vocal in her criticism of Trump, an online search for the words in the post returned no evidence she had ever said them.

Q7: Did Kamala Harris make the statement "And once Trump's gone and we have regained our rightful place in the White House, look out if you supported him and endorsed his actions, because we'll be coming for you next"?
A7: Chris Harris, the senator's communications director, also confirmed the quote is not real. Kamala Harris is one of several women still in consideration to be Joe Biden's running mate in the 2020 presidential election, The Associated Press has reported.

Q8: What did Kamala Harris allegedly say about Trump supporters at a rally?
A8: The bogus Facebook posts claim that Harris, a 2020 presidential candidate, said at a rally that she will seek "vengeance" against supporters of President Donald Trump.

Q9: Did Kamala Harris issue a threat to supporters of President Donald Trump?
A9: An image shared widely on Facebook claims that U.S. Sen. Kamala Harris issued a bold threat to supporters of President Donald Trump.

Q10: Is the quote attributed to Kamala Harris a genuine statement or a fabrication?
A10: Reuters, "Fact check: Kamala Harris quote on 'vengeance of a nation' fabricated by satire website," June 25, 2020

Scores:

Hungarian METEOR	Ev^2R Q-only	Ev^2R Q+A	SemQA
0.166	0.00	0.00	0.253

Table 12: Example 120. Claim: *"Kamala Harris said "And once Trump's gone and we have regained our rightful place in the White House, look out if you supported him and endorsed his actions, because we'll be coming for you next".* All metrics show low score.

Example 10 (Biden statement)

Claim: Biden has pledged to defund the police

Gold QA

Q1: Does Joe Biden support defunding the police?
A1: NO he said: "Federal dollars should not go to departments that violate people's rights or turn to violence as a first resort, but I donǵ support defunding police."

Generated QA

Q1: Does Joe Biden's police reform plan involve defunding the police?
A1: In fact, Biden's police reform plan calls for $300 million in funding for community-oriented policing, a portion of which would go toward hiring more police, which has been sharply criticized by some progressive Democrats.

Q2: What specific reforms to policing did Biden's criminal justice plan propose?
A2: Biden unveiled a criminal justice plan last year that tackled reforms to policing. The proposal calls for an injection of $300 million into the Community Oriented Policing Services program, which authorized money for the hiring of more police officers and policing training.

Q3: Does Joe Biden support the idea of defunding the police?
A3: Biden's campaign rejected the phrase "defund the police" and called for more funding for police departments to implement policy changes.

Q4: Does Joe Biden support defunding the police, or does he support conditioning federal aid to police departments based on certain standards?
A4: Former Vice President Joe Biden says he doesn't support defunding police, but he does support conditioning federal aid on whether police meet "certain basic standards of decency and honorableness.

Q5: Does Joe Biden support defunding the police?
A5: "I do not support defunding police," Biden wrote in an op-ed for USA Today. "The better answer is to give police departments the resources they need to implement meaningful reforms, and to condition other federal dollars on completing those reforms.

Q6: Does Joe Biden support the slogan "defund the police"?
A6: Biden's opposition to defunding police departments puts him at odds with activists and some progressive lawmakers who have amplified the slogan "defund the police" during demonstrations that have erupted nationwide in the wake of George Floyd's death in Minneapolis.

Q7: Does Joe Biden support defunding the police?
A7: Presumptive Democratic nominee Joe Biden definitively declared "I do not support defunding police," in an op-ed Wednesday, as protesters around the country increase their calls for overhauling the criminal justice system and President Trump attempts to tie Biden to the "Defund the police" movement.

Q8: Does Joe Biden support reducing the budget for local police departments?
A8: Biden's campaign had said earlier Monday that he backs advocates' calls to increase spending on social programs separate from local police budgets, but he also wants more funding for police reforms such as body cameras and training on community policing approaches.

Q9: What did Joe Biden suggest doing with police funding in the context of the interview with Ady Barkan?
A9: In fact, Biden was responding to a question not about defunding the police but about shifting some funding to social service agencies: "But do we agree that we can redirect some of the funding?" progressive activist Ady Barkan asked in a July 8 interview.

Q10: Does the phrase "defund the police" refer to eliminating police departments entirely or revisiting their functions and shifting funding to other services?
A10: While some argue police departments should be eliminated entirely, as PolitiFact National detailed in a June 9, 2020 article, the use of the phrase "defund the police" more typically means to revisit the functions of police departments and shift funding toward, for instance, mental health and social services.

Scores:

Hungarian METEOR	Ev^{2}R Q-only	Ev^{2}R Q+A	SemQA
0.375	1	1	1

Table 13: Example 10. Claim: *"Biden has pledged to defund the police".* All metrics show high score, and indicate an agreement.

Example 121 (Russian vaccine)

Claim: Russia has successfully developed a vaccine for Covid-19 and it has passed all checks.

Gold QA

Q1: Who has developed this vaccine?
A1: 'Sputnik V' has been developed by Moscow-based Gamaleya Research Institute of Epidemiology and Microbiology.

Q2: When was it first registered?
A2: It was registered by the Russian health ministry on August 11 as the first registered Covid-19 vaccine in the market but The certificate mentions that "the vaccine cannot be used widely until 1 January 2021", presumably after larger clinical trials are completed.

Q3: What kind of vaccine is it?
A3: The Russian vaccine is an "adenovirus vector-based vaccine".The idea is to use the weakened common cold virus to stimulate an immune response and trigger the formation of antibodies against Covid-19. These anti-bodies are then ready to fight against Covid-19 in case one catches a coronavirus infection. Even the widely awaited Oxford University vaccine is exploring the same route.

Q4: Has it been thoroughly tested?
A4: No

Q5: Has it been tested for safety?
A5: No

Generated QA

Q1: Has the COVID-19 vaccine developed by Russia completed all necessary clinical trials?
A1: Russia has not yet published any scientific data from its first clinical trials. The WHO's list of vaccine candidates in human testing still lists the Gamaleya product as in Phase 1 trials.

Q2: Has the COVID-19 vaccine developed by Russia completed a phase-three trial before its approval?
A2: But the vaccine's approval by Russian regulators comes before the completion of a larger study involving thousands of people, known as a phase-three trial.

Q3: Has the COVID-19 vaccine developed by Russia been proven to be safe and effective?
A3: Russian President Vladimir Putin announced the approval of a coronavirus vaccine for use on Tuesday, claiming it as a "world first," amid continued concern and unanswered questions over its safety and effectiveness.

Q4: How long did it take for the clinical trials of the Russian vaccine to be completed?
A4: Clinical trials of the Russian vaccine have been completed in less than two months and phase three trials are set to begin shortly, despite the vaccine having already been registered.

Q5: Has the COVID-19 vaccine developed by Russia been proven to be safe and effective?
A5: Scientists and public health officials are skeptical about Russian President Vladimir Putin's claim that the country's potential vaccine for the coronavirus "works quite effectively," saying Tuesday that the vaccine still needs critical testing to determine whether it's safe and effective.

Q6: How long did it take for the clinical trials of the Russian vaccine to be completed?
A6: Clinical trials of the Russian vaccine have been completed in less than two months and phase three trials are set to begin Wednesday, despite the vaccine having already been registered.

Q7: Does Russia have sufficient data to support the safety of its newly approved COVID-19 vaccine?
A7: Russian President Vladimir Putin says his country has become the first in the world to formally approve a new COVID-19 vaccine, despite a lack of data to back up his claims that the drug is safe to administer to humans.

Q8: Has the COVID-19 vaccine developed by the Gamaleya Institute in Moscow been thoroughly tested?
A8: Despite having only been in clinical trials for less than two months, the vaccine developed by the Gamaleya Institute in Moscow is safe, Putin said at a televised cabinet meeting, noting that it has already been given to one of his daughters, according to Reuters and The Washington Post.

Q9: What stages of clinical trials has the Russian Covid-19 vaccine completed?
A9: Few details have been made available about the Russian vaccine so far, but its approval was based on only the first two phases of clinical trials, which are designed to make early assessments on whether a potential vaccine can induce an immune response, and whether it's safe to administer in humans.

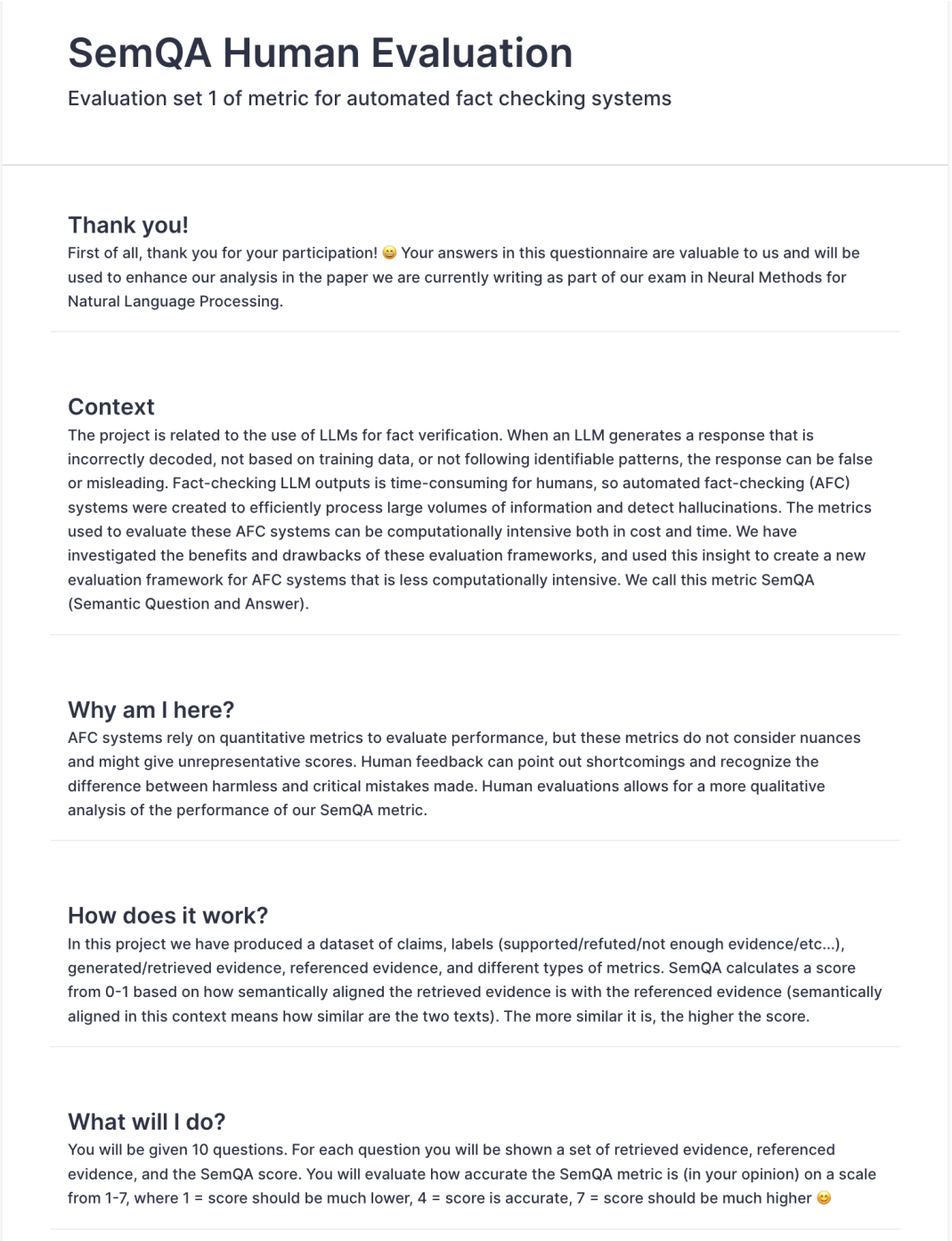
Q10: Has Russia's Covid-19 vaccine been proven to be safe and effective?
A10: Scientists are skeptical about Putin's claims on Russia's potential Covid-19 vaccine, saying the vaccine still needs critical data to determine whether it's safe or effective.

Scores:

Hungarian METEOR	Ev^{2}R Q-only	Ev^{2}R Q+A	SemQA
0.294	0.6	0.5	0.476

Table 14: Example 121. Claim: *"Russia has successfully developed a vaccine for Covid-19 and it has passed all checks."*. SemQA aligns closely with Ev^{2}r, while Hungarian METEOR is much lower.

B Human Evaluation form for SemQA Evaluation



SemQA Human Evaluation

Evaluation set 1 of metric for automated fact checking systems

Thank you!

First of all, thank you for your participation! 😊 Your answers in this questionnaire are valuable to us and will be used to enhance our analysis in the paper we are currently writing as part of our exam in Neural Methods for Natural Language Processing.

Context

The project is related to the use of LLMs for fact verification. When an LLM generates a response that is incorrectly decoded, not based on training data, or not following identifiable patterns, the response can be false or misleading. Fact-checking LLM outputs is time-consuming for humans, so automated fact-checking (AFC) systems were created to efficiently process large volumes of information and detect hallucinations. The metrics used to evaluate these AFC systems can be computationally intensive both in cost and time. We have investigated the benefits and drawbacks of these evaluation frameworks, and used this insight to create a new evaluation framework for AFC systems that is less computationally intensive. We call this metric SemQA (Semantic Question and Answer).

Why am I here?

AFC systems rely on quantitative metrics to evaluate performance, but these metrics do not consider nuances and might give unrepresentative scores. Human feedback can point out shortcomings and recognize the difference between harmless and critical mistakes made. Human evaluations allows for a more qualitative analysis of the performance of our SemQA metric.

How does it work?

In this project we have produced a dataset of claims, labels (supported/refuted/not enough evidence/etc...), generated/retrieved evidence, referenced evidence, and different types of metrics. SemQA calculates a score from 0-1 based on how semantically aligned the retrieved evidence is with the referenced evidence (semantically aligned in this context means how similar are the two texts). The more similar it is, the higher the score.

What will I do?

You will be given 10 questions. For each question you will be shown a set of retrieved evidence, referenced evidence, and the SemQA score. You will evaluate how accurate the SemQA metric is (in your opinion) on a scale from 1-7, where 1 = score should be much lower, 4 = score is accurate, 7 = score should be much higher 😊

Figure 3: Instructions to participants for human evaluations

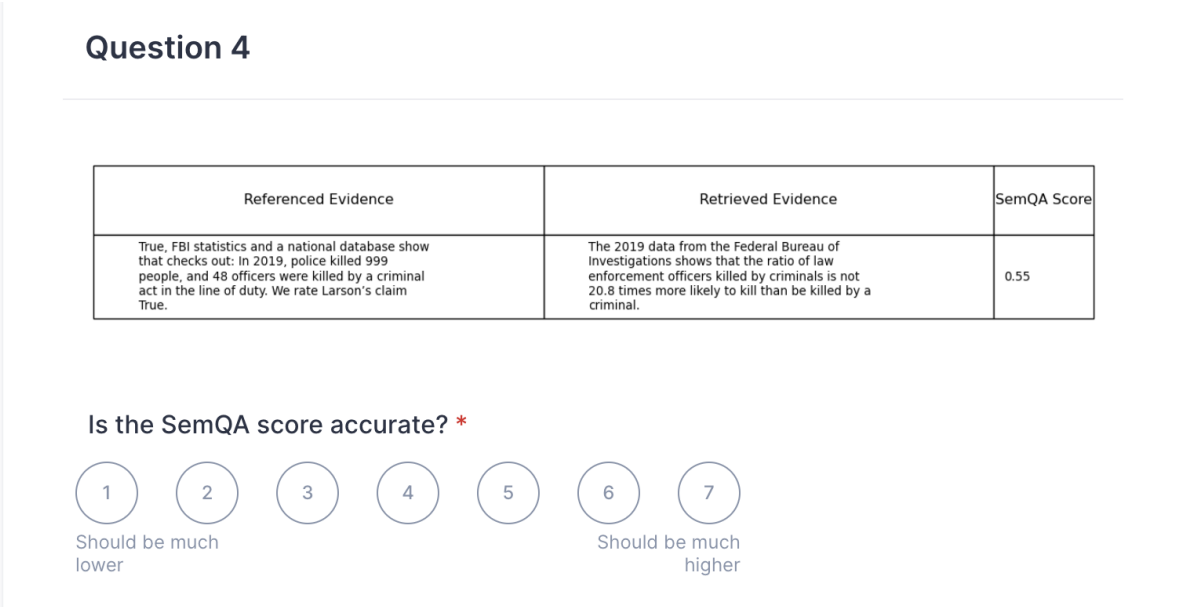


Figure 4: Example question in evaluation set for human evaluations

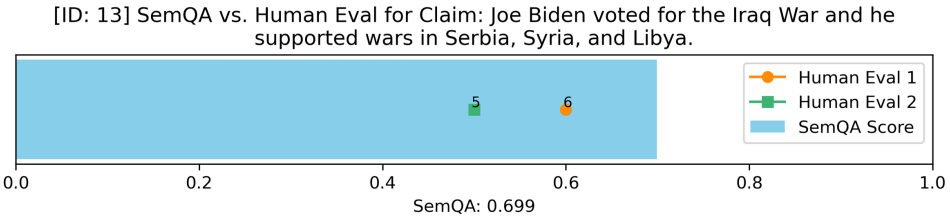


Figure 5: Comparison of SemQA score and human evaluations. Human evaluations of 5 and 6 means the SemQA score of 0.69 should have been slightly to moderately higher according to the annotators.

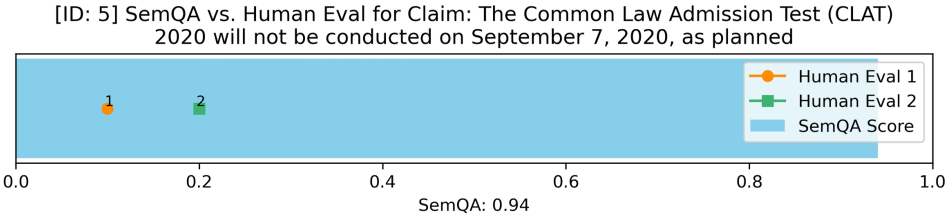


Figure 6: Comparison of SemQA score and human evaluations. Human evaluations of 1 and 2 means the SemQA score of 0.94 should have been far lower according to the annotators.

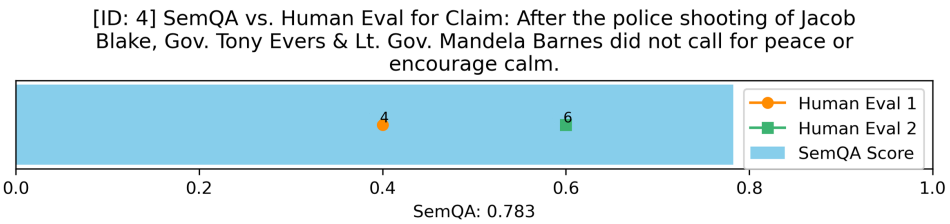


Figure 7: Comparison of SemQA score and human evaluations. Human evaluations of 4 and 6 means the annotators disagree; one marks the SemQA score of 0.78 as accurate, while the other marks the score as too low.