AVERIMATEC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web

Rui Cao[♡], **Zifeng Ding**[♡], **Zhijiang Guo**[♡], **Michael Schlichtkrull**[♢], **Andreas Vlachos**[♡]
University of Cambridge[♡], Queen Mary University of London[♢]
{rc990,zd320,zg283,av308}@cam.ac.uk, m.schlichtkrull@qmul.ac.uk

Abstract

Textual claims are often accompanied by images to enhance their credibility and spread on social media, but this also raises concerns about the spread of misinformation. Existing datasets for automated verification of image-text claims remain limited, as they often consist of synthetic claims and lack evidence annotations to capture the reasoning behind the verdict. In this work, we introduce **AVERIMATEC**, a dataset consisting of 1,297 real-world image-text claims. Each claim is annotated with question-answer (QA) pairs containing evidence from the web, reflecting a decomposed reasoning regarding the verdict. We mitigate common challenges in fact-checking datasets such as contextual dependence, temporal leakage, and evidence insufficiency, via claim normalization, temporally constrained evidence annotation, and a two-stage sufficiency check. We assess the consistency of the annotation in AVERIMATEC via inter-annotator studies, achieving a $\kappa = 0.742$ on verdicts and 74.7% consistency on QA pairs. We also propose a novel evaluation method for evidence retrieval and conduct extensive experiments to establish baselines for verifying image-text claims using open-web evidence.

1 Introduction

Misinformation has become a public concern due to its potential impact on elections, public health and safety [Allcott and Gentzkow, 2017, Aral and Eckles, 2019, Bär et al., 2023, Rocha et al., 2021]. To curb its spread, professional fact-checkers are employed to identify misleading content. However, they are unable to keep up with the vast volume of information online [Pennycook, 2019, Drolsbach et al., 2024]. The severity of the problem, along with the limitations of manual verification, has motivated the development of automated fact-checking (AFC) [Guo et al., 2022, Nakov et al., 2021].

To support research in AFC, the research community has created various benchmark datasets [Thorne et al., 2018, Aly et al., 2021, Schlichtkrull et al., 2023, Alhindi et al., 2018, Yao et al., 2023, Chen et al., 2024], aiming to enhance the effectiveness and interpretability of fact-checking systems. However, most existing benchmarks focus exclusively on textual claims, overlooking the important role of media in the dissemination of misinformation. Recent studies estimate that approximately 80% of online claims are multimodal involving both text and media [Dufour et al., 2024], as media can enhance perceived credibility [Newman et al., 2012] and increase exposure [Li and Xie, 2020]. Among these, images are the most prevalent media type [Dufour et al., 2024].

While several datasets have been developed for image-text AFC, many are synthetic, generated by manually manipulating either the textual or visual modality of image-text pairs [Luo et al., 2021, Papadopoulos et al., 2024, Jia et al., 2023]. Due to discrepancies between synthetic data and real-world data [Zeng et al., 2024], models that perform well on synthetic benchmarks may fail to generalize to real-world claims. Moreover, recent work [Papadopoulos et al., 2025] showed that models can achieve high performance on such datasets by exploiting superficial correlations, such as image-text similarity, without examining factuality and logical consistency. Some benchmarks [Zlatkova et al.,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/085713-0154

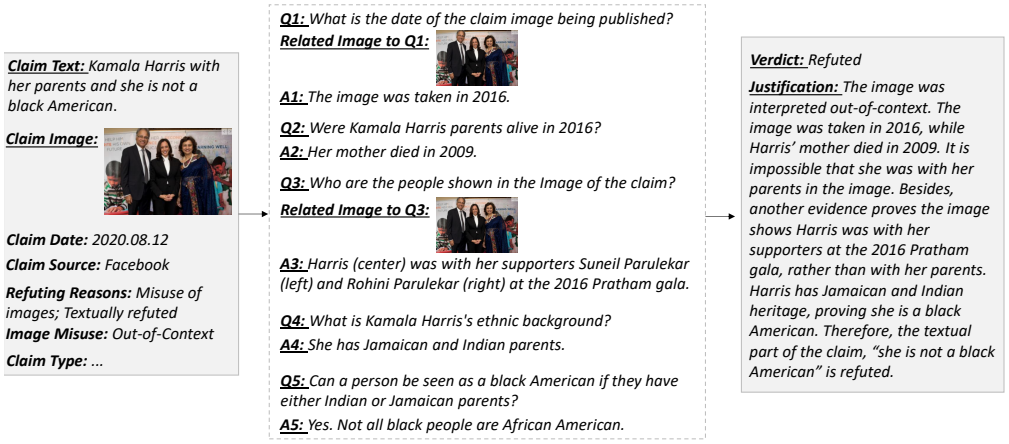

Figure 1: **An annotated claim from AVERIMATEC.** The rationale for verifying an image-text claim has been decomposed into a sequence of QA pairs, which could be potentially multimodal.

2019, Nakamura et al., 2020, Shu et al., 2020] attempt to include real-world image-text claims extracted from fact-checking articles. As noted in prior work [Ousidhoum et al., 2022, Schlichtkrull et al., 2023], this may result in omitting critical contextual information for verification, such as context to resolve coreferences. Additionally, both synthetic and real-world image-text datasets typically lack annotated evidence, making it impossible to evaluate models' reasoning process.

To address the limitations above, we propose the **A**utomated **Ver**ification of **Ima**ge-**Te**xt **C**laim (**AVERIMATEC**) dataset, where the verification of real-world image-text claims has been decomposed into a sequence of question-answering with evidence from the web. In addition, each claim is annotated with metadata information, a veracity label based on retrieved evidence and a textual justification, explaining how the verdict is reached, as shown in the example in Figure 1.

To construct AVERIMATEC, initially annotators are asked to identify and normalize image-text claims from fact-checking articles, incorporating necessary contextual information while providing associated metadata. Next, annotators convert the verification rationale from the articles into QA pairs, while being restricted to using only online evidence published before the claim's date. Given the multimodal nature of the task, both questions and answers may involve images. Finally, we conduct two rounds of quality control to ensure that each annotated claim is supported by sufficient evidence for the annotated verdict. The resulting dataset, contains 1,297 image-text claims. To assess the consistency of the verdict labels we conducted an inter-annotator agreement study in which we obtained a Randolph's [Randolph, 2010] free-marginal κ of 0.742 over 100 re-annotated claims. The re-annotation recovered 74.7% of the original QA pairs, confirming that the annotations capture reasoning paths for verifying image-text claims consistently.

We further introduce a baseline for image-text claim verification, which operates by generating evidence-seeking questions aimed at fact-checking and answering these questions with a set of expert tools. Since AVERIMATEC is the first image-text claim verification dataset to incorporate QA annotations that explicitly reflect reasoning paths and evidence, we develop a reference-based evaluation method to assess models' generated questions and retrieved evidence. Using this evidence evaluation, we report conditional verdict accuracy which measures the correctness of predicted verdicts only when the associated evidence score exceeds a predefined threshold[1].

2 Related Works

Automated fact-checking (AFC) has become increasingly important due to the pressing need to curb the spread of misinformation [Guo et al., 2022, Nakov et al., 2021]. To support AFC research, several datasets focusing primarily on text-based claims have been proposed (see the top block of

[1]The dataset is available here: https://huggingface.co/datasets/Rui4416/AVerImaTeC. The code can be accessed here: https://github.com/abril4416/AVerImaTeC

Table 1: **Comparison of fact-checking datasets.** *Indep.* (independence) denotes whether extracted claims are context independent (e.g., understandable without fact-checking articles). *Img.*, *Suff.*, *Retr.*, and *Unleak.* are abbreviations for image, sufficiency, retrieval and unleaked evidence, respectively. Sufficiency indicates whether there is sufficient supporting evidence to reach annotated verdicts; Retrieval refers to whether open-world evidence retrieval is performed; Unleaked represents whether annotated evidence contains temporal leakage such as including evidence published after claims.

Dataset	Claim		Evidence				# Claims
	Real	*Indep.*	*Img.*	*Suff.*	*Retr.*	*Unleak.*	
FEVER [Thorne et al., 2018]	✘	✓	✘	✓	✓	-	185,445
FEVEROUS [Aly et al., 2021]	✘	✓	✘	✓	✓	-	87,026
Liar-Plus [Alhindi et al., 2018]	✓	✘	✘	✓	✘	✘	12,836
Snopes [Hanselowski et al., 2019]	✓	✘	✘	✘	✘	✓	6,422
MultiFC [Augenstein et al., 2019]	✓	✘	✘	✘	✓	✘	36,534
AVeriTec [Schlichtkrull et al., 2023]	✓	✓	✘	✓	✓	✓	4,568
CLAIMDECOMP [Chen et al., 2024]	✓	✘	✘	✘	✓	✓	1,200
MOCHEG [Yao et al., 2023]	✓	✘	✓	✘	✘	✘	15,601
NewsCLIPpings [Luo et al., 2021]	✘	✓	-	-	-	-	988,283
InfoSurgeon [Fung et al., 2021]	✘	✓	-	-	-	-	30,000
Autosplice [Jia et al., 2023]	✘	✓	-	-	-	-	5,894
DGM [Shao et al., 2023]	✘	✓	-	-	-	-	230,000
MMFake [Liu et al., 2024b]	✘	✓	-	-	-	-	11,000
Verite [Papadopoulos et al., 2024]	✘	✓	-	-	-	-	1,000
COSMOS [Aneja et al., 2021]	Mix	✘	-	-	-	-	201,700
FACTIFY [Mishra et al., 2022]	Mix	✘	-	-	-	-	50,000
FACTIFY 2 [Suryavardan et al., 2023]	Mix	✘	-	-	-	-	50,000
MMOOC [Xu et al., 2024]	Mix	✓	-	-	-	-	364,000
Fauxtography [Zlatkova et al., 2019]	✓	✘	-	-	-	-	1,233
Fakeddit [Nakamura et al., 2020]	✓	✘	-	-	-	-	1,063,106
Qprop [Barrón-Cedeño et al., 2019]	✓	✘	-	-	-	-	51,294
FakeNewsNet [Shu et al., 2020]	✓	✘	-	-	-	-	23,196
MuMiN [Nielsen and McConville, 2022]	✓	✘	-	-	-	-	12,914
AVERIMATEC	✓	✓	✓	✓	✓	✓	1,297

Table 1). Motivated by the prevalence of images in claims, fact-checking datasets for image-text claims were introduced (listed in the bottom block of Table 1). Some studies [Luo et al., 2021, Papadopoulos et al., 2024, Jia et al., 2023] have generated synthetic claims by applying manipulation techniques to the visual and textual modalities of image-text pairs. However, there are discrepancies between synthetic data and real-world image-text claims [Zeng et al., 2024, Papadopoulos et al., 2025], raising concerns about the generalization of models to real-world image-text claim verification. Although some benchmarks [Zlatkova et al., 2019, Barrón-Cedeño et al., 2019, Shu et al., 2020] focus on real-world claims (e.g., those derived from fact-checking articles), they all suffer from context dependence. Moreover, all existing datasets with image-text claims lack evidence annotations, limiting transparency, and the ability to understand the rationale behind fact-checking verdicts.

To capture the rationale in claim verification, a complex reasoning task, various reasoning representations have been explored, including natural logic [Strong et al., 2024], functional programs [Pan et al., 2023b], and QA [Chen et al., 2024, Qi et al., 2023, Pan et al., 2023a]. In AVerImaTeC, we adopt QA as the reasoning representation, as the QA format is more intuitive and accessible for annotators, enabling efficient and consistent annotation by non-experts. Furthermore, other reasoning forms can often be mapped into QA-style representations, ensuring compatibility and flexibility for future extensions.

3 Annotation Schema

Each claim in AVERIMATEC is normalised to be understandable alone without additional context, such as the original social media post or the associated fact-checking article. For each claim, we provide key metadata such as the *speaker*, *publisher*, *publication date*, and the relevant *location*.

https://doi.org/10.52202/085713-0154

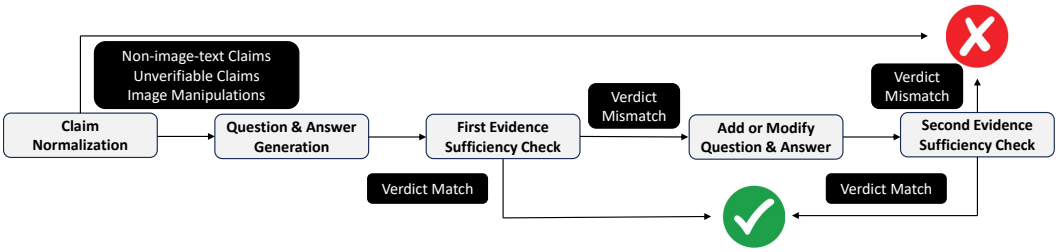

Figure 2: **Annotation pipeline.** We first normalize the claim, then perform QA annotation to structure evidence retrieval. Two rounds of evidence sufficiency checks ensure annotation quality.

These metadata elements can serve as valuable evidence for claim verification. We further annotate each claim with *claim type* (e.g., *quote verification*, which determines whether a quote was actually attributed to the correct speaker), and *fact-checking strategy* (e.g., *reverse image search* to find background information about images). While not all of these annotations are directly used during verification, they offer insights for developing fact-checking models.

The sequence of QA pairs reflects the reasoning process involved in evidence retrieval and verification. Given the multimodal nature of the claims, both questions and answers may include images (i.e., image-related questions and image answers). A single question may have multiple answers, as there may be conflicts or disagreements in the evidence. Questions may refer to previous questions and their answers as long as they are understandable on their own, thus capturing multi-hop reasoning in fact-checking. Answers (other than *"No answer could be found."* or derived via image analysis without external information) must be supported by a *source url* linking to a web page. To ensure long-term accessibility, all source pages are archived on the internet archive.[2] We also provide metadata for each annotated QA pair, including the *question type* (e.g., *image-related* or *metadata-related*), *answering method*, *answer type* (e.g., *extractive*, *abstractive*, *boolean* or *image-based*), and *source medium type* (type of web content used as evidence).

We follow the four-way veracity labeling schema from Schlichtkrull et al. [2023]: *supported*, *refuted*, *not enough evidence*, and *conflicting/cherry-picking*. An image-text claim can be refuted due to its textual part (the textual part of is factually wrong) and/or due to *misuse of images* (e.g., misinterpreting the context of an image). For instance, the claim in Figure 1 is refuted due to both textual refutation and image misuse. *Not enough evidence* refers to cases where evidence is insufficient to either support or refute a claim. *Conflicting/Cherry-picking* covers cherry-picking claims, true-but-misleading claims, as well as claims with conflicting evidence. Conflicts among evidence has been extensively studied in the context of QA [Liu et al., 2024a, Lee et al., 2024].

A textual justification is added to the claims to explain how verdicts can be reached on the basis of the evidence found. Considering that the evidence can contain multiple images, we assign unique and special tokens to the images in the evidence (e.g., `[IMG_1]`) for annotators to refer to in justifications. They may also include commonsense reasoning or inductive reasoning beyond the retrieved evidence. For instance, the evidence of the claim in Figure 1 proves the image was taken in 2016 while Kamala Harris' mother died in 2009. The justification infers it is impossible that Harris *was with her parents in the image*, taken in 2016 as her mother died in 2009 already.

4 Annotation Process

The five-phase annotation pipeline is illustrated in Figure 2, extending the annotation process proposed in Schlichtkrull et al. [2023] to the domain of image-text claims. In phase one, an annotator extracts and normalize valid image-text claims, given a fact-checking article. For each extracted claim, a different annotator in the second phase generates questions and answers to reflect the rationale of fact-checking based on the article, using evidence from the web. A provisional veracity label is annotated as well. Thirdly, a first round of evidence sufficiency checking will be conducted by a third annotator who provides a justification and a verdict solely based on the QA pairs, without considering the fact-checking article. Different verdicts in the second and third phases suggest possible insufficient

[2]https://archive.org/

Table 2: **Data statistics for dataset splits.** End date refers to the latest publication date of claims included in each split. The start date of each *dev* and *test* split corresponds to the end date of the preceding split. The final row of the table reports the distribution of claim labels across four categories: *supported* (S), *refuted* (R), *conflicting/cherry-picking* (C), and *not enough evidence* (N).

Split	Train	Dev	Test
# Claims	793	152	352
# Images / Claim	1.49	1.38	1.38
# QA Pairs / Claim	2.86	2.84	3.11
Reannotated (%)	15.0	15.8	9.4
End Date	31-05-2023	31-07-2023	21-03-2025
Labels (S / R / C / N) (%)	1.6 / 95.3 / 0.8 / 2.3	2.6 / 92.8 / 0.7 / 3.9	13.9 / 78.1 / 2.0 / 6.0

evidence in the second phase. In this case, the claim is forwarded to a fourth phase to add or modify existing QA pairs and a fifth phase for an additional round of evidence sufficiency check. Any claims for with unresolved conflicts in the verdicts are discarded. We ensure that each annotation phase of a claim is done independently by using different annotators (i.e., the same claim will not be annotated by the same annotator twice). Details for annotation guidelines and annotators' demographics are provided in Appendix B and C, respectively.

Claim Extraction & Normalization. Given a fact-checking article, an annotator extracts all claims from it, as multiple claims may pertain to a single event. The modality types of claims are annotated, and only *image-text* claims (i.e., textual claims paired with images) are forwarded to phase two. For identified image-text claims, sufficient context must be provided to ensure that the claim, specifically its textual component, can be understood independently of the surrounding article. In some cases, metadata (e.g., date, location, speaker) is sufficient to disambiguate a claim. However, for ambiguities beyond metadata, such as unresolved coreference or missing referents, annotators are instructed to enrich the extracted claims so that they can be interpreted independently. The associated images are uploaded and normalized (e.g., separating collages or locating original, unaltered versions). Relevant metadata is also annotated. We exclude unverifiable claims (e.g., speculation or personal opinions) and claims where images are not used in verification or involve manipulated content.

Question Generation & Answering. Annotators in this phase are instructed to transform the rationales derived from fact-checking articles into a sequence of QA pairs. Each question is annotated with its *question type*, *answering method* and *answers*. For *image-related* questions, annotators must select relevant images from either the claim images or images from previous answers. If a question is not marked as *unanswerable* or pertains to *image analysis*, annotators are required to provide supporting urls of the evidence source. We advise annotators to prioritize evidence sources linked within the fact-checking articles but to exclude anything published after the claim date, including the article itself. When linked sources are unavailable (e.g., dead links) or insufficient, annotators are provided with a custom Google search interface that supports both text and image queries. All retrieved pages from the interface are restricted to dates *prior* to the claim date to prevent temporal leakage [Glockner et al., 2022]. Based on the generated QA pairs, annotators assign a verdict.

Evidence Sufficiency Check. In this phase, a third annotator, who does not have access to the fact-checking article, is presented with the extracted claim and its associated annotated QA pairs. The annotator is tasked with assigning a verdict and providing a textual justification. This verdict is then compared to the one generated during the QA annotation phase. A discrepancy between the two verdicts indicates insufficient evidence, and the QA annotation process is repeated to refine the QA pairs, followed by a second round of evidence sufficiency assessment with new annotators.

5 Dataset Statistics

Data Distribution. We began with 2,353 fact-checking articles After discarding those that were inaccessible, did not focus on image-text claims, or contained unresolved annotation conflicts, we obtained 1,297 annotated image-text claims using the annotation pipeline described in the previous section. Further details on article sources and the filtering process are provided in Appendix D.1 and D.2, respectively. The splits of our dataset are temporally organized, and detailed statistics are presented in Table 2. We find that 23.7% image-text claims include more than one claim image,

https://doi.org/10.52202/085713-0154

highlighting the need to understand multiple visual inputs. On average, each claim is annotated with 2.92 questions. Of these, 3.5% questions have more than one answer, and 62.5% are image-related, emphasizing the importance of visual context in verifying image-text claims. To address these questions, annotators frequently selected *Image-search* (53.9%) as the answering method, indicating the necessity for tools that support image-centric information retrieval. Regarding answer types, 58.8% are *extractive*, consistent with our annotation guidelines. Additionally, 1.6% of answers are images themselves, underscoring the importance of supporting image retrieval as direct answers. A small proportion (2.6%) of questions are marked as *unanswerable*, reflecting cases where no supporting evidence could be found online. Further metadata statistics, such as *claim types* and *answer types*, are provided in Appendix D.3. The dataset shows a label imbalance, with most claims being *refuted*, which is expected given that misleading content is more likely to be scrutinized.

Inter-Annotator Agreement. We re-annotated 100 claims with a different group of annotators, following the same annotation pipeline as described in Section 4. As in prior work [Schlichtkrull et al., 2023], we assume that the first phase of annotation has already been completed, and thus the re-annotation process begins from the second phase. We evaluate inter-annotator agreement for both verdict labels and QA annotations. For verdict agreement, we use Randolph's [Randolph, 2010] free-marginal multi-rater κ, designed for unbalanced datasets [Warrens, 2010]. We obtained an agreement score of $\kappa = \mathbf{0.742}$. For comparison, AVeriTec [Schlichtkrull et al., 2023] reported agreement of 0.619. Using Fleiss' κ, a more traditional metric, our annotation process achieves an agreement score of 0.450. To evaluate QA annotations, the three best performing annotators were provided with an extracted claim and two independently annotated sets of QA pairs, and asked to determine how many QA pairs in one set were covered by the other. We compute recall and precision by comparing the original annotated QA pairs against those from re-annotation. The recall rate is $\mathbf{74.7}\%$ and the precision is $\mathbf{67.2}\%$. The substantial overlap between the two sets suggests strong agreement between annotators.

6 Evaluation

The evaluation of model accuracy on our dataset considers both the *retrieved evidence* and the *veracity*. Following [Thorne et al., 2018, Schlichtkrull et al., 2023], we first assess the quality of the retrieved evidence by comparing it against human-annotated references, and report veracity prediction accuracy conditioned on the evidence scores. I.e., the accuracy of a verdict prediction is considered only if the associated evidence score exceeds a predefined threshold λ, otherwise the claim is considered to be labeled incorrectly. This reflects the requirement that an effective fact-checking system should not only predict return verdicts on claims but also provide appropriate evidence.

Recent research [Akhtar et al., 2024] showed that reference-based evaluation of evidence with large language models (LLMs) aligns best with human assessments. We extend this framework to a multimodal setting where evidence from the web may comprise text and images. We transform each QA pair into an evidence statement following [Akhtar et al., 2024], where images are represented with special image tokens (e.g., `[IMG_1]`). For instance, the first QA pair in Figure 1 is transformed to *"[IMG_1] was published in 2016."*. We then conduct separate reference-based evaluations for the textual and visual components. For the textual part, we use a validation method similar to [Akhtar et al., 2024]. If a matched evidence item is found within the ground-truth set, we proceed to a second step to compare the associated images. If image similarity falls below a threshold the match is deemed invalid due to image mismatch. We exploit Gemini-2.0-Flash [DeepMind, 2024] as the scoring model for both steps in the reference-based evaluation inspired by its power [Akhtar et al., 2024]. We report evidence *recall*, defined as the percentage of ground-truth evidence instances successfully retrieved.

We performed alignment checks and robustness checks (against adversarial attacks) of different reference-based evaluation schemes (text-only, interleaved and the separated evaluation) in our setting (details in Appendix F.2 and F.3). Our separated reference-based evaluation achieved the highest alignment with human assessments, with a Spearman correlation coefficient (ρ) [Spearman, 1904] of 0.332 and a Pearson correlation coefficient (r) [Pearson and Henrici, 1896] of 0.381. Furthermore, the separated evaluation method is the most robust towards adversarial attacks.

Our evaluation primarily focuses on evidence retrieval and verdict prediction. Additionally, given the availability of justification annotations, we assess the quality of model-generated justifications by comparing them to the human-annotated ground truth with the traditional evaluation metric,

Table 3: **Experimental results of baselines on AVERIMATEC. Q-Eval** and **Evid-Eval** denote for *recall* scores of generated questions and retrieved evidence, with reference of ground-truth questions and evidence. We report verdict prediction and justification generation scores conditioned on evidence retrieval performance, specifically only considering verdict accuracy and justification generation performance when the evidence score is above 0.2, 0.3 and 0.4.

LLM	**MLLM**	**Q-Eval**	**Evid-Eval**	**Veracity** (.2/.3/.4)			**Justifications** (.2/.3/.4)		
Paralleled Question Generation									
Gemini	Gemini	0.42	0.15	0.15	0.13	0.08	0.15	0.11	0.07
Qwen	Qwen-VL	0.43	0.18	0.09	0.08	0.05	0.13	0.11	0.07
Gemma	Gemma	0.39	0.21	0.14	0.12	0.09	0.17	0.14	0.10
Qwen	LLaVA	0.37	0.16	0.09	0.08	0.05	0.12	0.10	0.06
Dynamic Question Generation									
Gemini	Gemini	0.33	0.22	0.17	0.16	0.12	0.17	0.16	0.11
Qwen	Qwen-VL	0.27	0.12	0.10	0.09	0.05	0.09	0.08	0.05
Gemma	Gemma	0.27	0.19	0.15	0.13	0.10	0.15	0.13	0.09
Qwen	LLaVA	0.32	0.16	0.13	0.11	0.08	0.11	0.10	0.08
Hybrid Question Generation									
Gemini	Gemini	0.36	0.19	0.18	0.17	0.10	0.17	0.15	0.09
Qwen	Qwen-VL	0.37	0.16	0.11	0.09	0.06	0.12	0.10	0.06
Gemma	Gemma	0.26	0.25	0.16	0.15	0.11	0.19	0.17	0.12
Qwen	LLaVA	0.30	0.17	0.09	0.09	0.07	0.12	0.11	0.07

ROUGE-1 [Lin, 2004], a standard metric that is relatively tolerant of longer outputs. While ROUGE-1 provides a baseline assessment, we recognize that more sophisticated and targeted evaluation methods may be necessary. Similarly, we adopt a conditional justification generation score that any claim for which the evidence score is below λ receives a score for justification of 0. The prompts used for the QA pair conversion and evaluation are provided in Appendix I.1 and I.2, respectively.

7 Experiments

7.1 Baselines

Our baseline system guides the fact-checking process by sequentially posing and answering essential questions to verify image-text claims. It consists of four components: a *question generator*, an *answer generator*, a *verifier* and a *justification generator*. Throughout the question-answering process, the system maintains an evolving *evidence history context*, which is continuously updated with summarized evidence derived from each QA pair. After the QA stage, the verifier receives the claim and the accumulated evidence context to predict a veracity label. Finally, the justification generator produces a rationale that explains how the predicted verdict is supported by the evidence.

Question Generator. A straightforward approach is to generate all verification questions at once, given an image-text claim, as in prior work [Schlichtkrull et al., 2023]. We refer to this strategy as *paralleled* question generation (**PQG**). However, since later questions often depend on earlier ones (e.g., as shown in Figure 1), we also propose a *dynamic* question generation (**DQG**) method, where each subsequent question is generated based on both the claim and the evolving evidence history. To combine the strengths of both approaches, we introduce a *hybrid* question generation (**HQA**) method, which generates the first few questions in parallel and then switches to dynamic generation for the remaining ones. All three strategies employ an MLLM for question generation, leveraging the model's internal decomposition capability. The prompts used for each strategy are detailed in Appendix I.3.

Answer Generator. Given a generated question, the answer generator is responsible for generating an answer to the question. Inspired by recent research on tool usage [Wu et al., 2023, Cao et al., 2024, Braun et al., 2024], we integrate a set of specialized tools into the answer generation module, along with a tool selector that automatically selects the appropriate tool for a given question. The system includes tools for: (1) *reverse image search* (RIS) to retrieve text-based information associated with

https://doi.org/10.52202/085713-0154

Table 4: **Baseline performance on verdict prediction and justification generation with ground-truth evidence (first block) and without accessing any external evidence (second block).** NEE is for *Not Enough Evidence* and Conflict. is for *Conflicting/Cherry-picking*. Justi. is for the performance of justification generation.

Evid. Source	LLM	MLLM	Refuted	Supported	NEE	Conflict.	Overall	Justi.
Ground-Truth	Gemini	Gemini	0.84	0.92	0.52	0.00	0.82	0.50
	Qwen	Qwen-VL	0.87	0.47	0.62	0.00	0.78	0.44
	Gemma	Gemma	0.63	0.84	0.62	0.00	0.64	0.49
	Qwen	LLaVA	0.55	0.47	0.90	0.00	0.55	0.43
No Search	Qwen	Qwen-VL	0.01	0.02	0.14	0.00	0.02	0.04

an image; (2) *web search for texts* (WST) to retrieve relevant web texts given a textual query; (3) *web search for images* (WSI) to retrieve relevant images using a textual query; and (4) *visual question answering* (VQA) to answer questions directly based on input images, including comparison and detail analysis. Among these, VQA, implemented with an MLLM, directly outputs an answer, while RIS and WST are followed by an LLM to leverage the retrieved text to generate an answer. WSI is followed by an MLLM, which incorporates the retrieved image into the answer generation process. Prompts for tool selection and answer generation are provided in Appendix I.4 and I.5, respectively.

Verifier. The verifier takes the claim to be verified, the multimodal evidence context and the verdict definitions as introduced in Section 3 to predict a veracity label for the claim (prompts in use shown in Appendix I.6). An MLLM will serve as the verifier.

Justification Generator. Given the predicted verdict, in this step, the justification generator is asked to provide an explanation for the prediction. It also has access to the claim and evidence history (detailed prompts in Appendix I.7). We rely on an MLLM for justification generation.

Model Implementation. Our baseline system includes both an LLM and an MLLM, which could take different roles in components. We experimented with four combinations of LLMs and MLLMs: 1) Gemini-2.0-flash-001 [DeepMind, 2024] (**Gemini**) as both the LLM and the MLLM; 2) Qwen/Qwen2.5-7B-Instruct [Yang et al., 2024] (**Qwen**) acts as the LLM and Qwen2.5-VL-7B-Instruct [Bai et al., 2025] (**Qwen-VL**) serves as the MLLM; 3) Gemma-3-12B [Kamath et al., 2025] (**Gemma**), capable of both unimodal and multimodal understanding, is used as both the LLM and the MLLM; and 4) Qwen and LLaVA-Next-7B [Li et al., 2024] (**LLaVA**) work as the LLM and MLLM respectively. More details on model implementation and experiment settings are in Appendix H.

7.2 Main Results

We investigated the performance of baseline models both under zero-shot and few-shot settings. In the latter, three training instances were used as demonstrations to guide question generation (details in Appendix H.2). Overall, few-shot baselines slightly outperform their zero-shot counterparts. Few-shot performance results for various combinations of LLMs and MLLMs are presented in Table 3 (zero-shot model performance in Appendix G.1, and the findings hold for both settings). We report conditional veracity accuracy and ROUGE-1 scores for justification generation under varying evidence evaluation thresholds, $\lambda = \{0.2, 0.3, 0.4\}$.

Comparison of Question Generation Strategies. Among the three question generation strategies, the parallel approach consistently outperformed others in producing critical questions for fact-checking. The dynamic strategy, where questions are generated based on evolving evidence, yielded a weaker performance. This suggests that the increased complexity of reasoning over evolving interleaved image-text evidence poses significant challenges for current MLLMs. Open-source MLLMs often generated repetitive questions when using the dynamic strategy, further highlighting their limitations in handling complex multimodal inputs.

Performance of Evidence Retrieval. Evidence scores across all models are significantly lower than their question evaluation scores, underscoring the difficulty of retrieving appropriate evidence for image-text verification. One reason for the failure of evidence retrieval is that models exhibited a bias toward using VQA as the answering tool (e.g., Qwen + Qwen-VL in PQG selected VQA as the answering tool for 30% questions), diverging from human fact-checkers' preferences, despite being

provided with tool selection demonstration examples (more elaborations in Appendix H). MLLMs tend to rely more on internal image details rather than external contextual information to address image-related questions. Additionally, approximately 13% images failed to retrieve any contextual information via RIS (i.e., no web pages published before claim dates could be found), consistent with the findings of [Tonglet et al., 2024]. A notable portion of questions, e.g., 30% questions of the Qwen + Qwen-VL baseline with PQG, elicited responses such as *"No answer could be found"*, based on retrieved evidence. This can be attributed to two main factors besides the failure of RIS: (1) many web pages retrieved by RIS were non-scrapable (e.g., Instagram posts), and (2) the baseline evidence employed a naive ranking method, BM25 [Robertson and Zaragoza, 2009], which neither considered the visual content of images nor incorporated fine-grained re-ranking. Interestingly, higher scores for question generation did not always translate into better evidence retrieval. For instance, under the hybrid strategy, the Gemini-based baseline model achieved a 0.1-point higher question generation score than the Gemma-based model, but had worse evidence retrieval performance. Further analysis showed that Gemini generated a higher proportion of RIS-dependent questions (40.1% vs. 32.4%), and the majority (62.6% vs. 31.6%) of these were unanswerable using the retrieved evidence.

7.3 Analysis and Discussion

We conducted analysis to assess baseline model performance under two conditions: (1) using ground-truth evidence (first block of Table 4), and (2) disabling web-based evidence retrieval (second block of Table 4). For both settings, we report conditional accuracy and justification generation scores with the evidence threshold set to $\lambda = 0.3$.

Baselines' Performance with Ground-truth Evidence. The results obtained using golden evidence represent upper-bound performance, highlighting the models' full potential. Most baselines perform well in verdict prediction under this setting, underscoring both the importance and difficulty of effective evidence retrieval. Notably, the Gemini-based baseline achieves the highest scores for both prediction and justification generation. In contrast, baselines using LLaVA as the MLLM demonstrate the weakest performance. This is expected, as LLaVA [Li et al., 2024] was not pre-trained on interleaved image-text documents. Across all baselines, we observe consistent failure in identifying conflicting claims. This result aligns with prior findings on the challenge of detecting conflicting textual claims [Schlichtkrull et al., 2023]. Moreover, our dataset contains a relatively small number of conflicting claims, which can cause instability in model performance. It is also worth noting that our models do not explicitly model conflict within evidence, in contrast to [Schlichtkrull et al., 2023], which predicted verdicts per evidence piece and then examined whether these verdicts conflicted. In our setup, we observe stronger dependencies between multiple QA pairs that individual QA pairs are often insufficient to support or refute a claim. For example, only by combining the first and second QA pairs in Figure 1 can the model conclusively refute the claim.

Although baselines used ground-truth evidence, they still achieved low justification generation scores. We attribute this to the limitations of ROUGE-1 for evaluating justification generation. To address this, we experimented with a reference-based evaluation method, Ev2R [Akhtar et al., 2024], which has been shown to align well with human assessments in open-ended generation tasks. Using this approach, the justification scores were much higher and more encouraging (details in Appendix G.2). Moving forward, we plan to explore and incorporate more appropriate evaluation methods for justification generation.

Baselines' Performance without Searching. For the baseline without external evidence retrieval, we selected the combination of Qwen + Qwen-VL under PQG, as it demonstrated the strongest performance in question generation for claim verification. In this setting, all questions that require external search, those invoking the RIS, WST, or WSI tools, are marked with the response *"No answer could be found."*. As shown in the second block of Table 4, the model achieves an evidence score of 0.06, reflecting a significant drop in evidence retrieval. This result underscores the critical role of web-based information retrieval in real-world claim verification tasks. Nonetheless, in a few isolated cases, the model was still able to generate accurate evidence. These instances typically involved questions about locations or events depicted in the image, which the VQA tool could address correctly, likely because the model encountered these images during pretraining. However, such behavior also raises concerns about potential data leakage from MLLMs. Since the model's responses are not grounded in externally retrieved evidence, there is a risk that its predictions stem from memorized content or prior exposure to fact-checking articles during training.

https://doi.org/10.52202/085713-0154

8 Limitations

AVerImaTeC has a relatively limited scale, as it is sourced from real-world claims and constructed through detailed human annotation. This is consistent with other human-annotated datasets of real-world claims, such as AVeriTeC [Schlichtkrull et al., 2023], which also contain only a few thousand claims. Since the claims in AVerImaTeC originate from fact-checking articles, the dataset may inherit biases inherent in these sources—for example, selection bias [Shin and Thorson, 2017, Barnoy and Reich, 2019], leading to imbalanced label distributions.

We made considerable efforts to prevent temporal leakage. Specifically, we provided annotators with custom search bars to retrieve online evidence published prior to the claim date. However, the exact dates of a small portion of claims (5%) were unavailable. In such cases, we used the dates of the corresponding fact-checking articles, which may have been published a few days after the original claims. Additionally, we relied on Google Search and the Python package *htmldate.find_date* to estimate publication dates of web pages, which are coarse approximations.

Furthermore, we exploited a reference-based evaluation strategy for both the generated questions and the retrieved evidence. Though it aligns well with human assessments, it has limitations in cases where model predictions are reasonable but not reflected in the reference annotations. In such cases, evaluation scores may be undeservedly low due to poor alignment with the references.

9 Ethical Statement

The datasets and models described in this paper are not intended for use in truth-telling tasks, such as automated content moderation systems. The labels and justifications included in the dataset reflect only the evidence recovered by annotators and are therefore subject to the biases of both annotators and journalists. In addition, the QA annotations may introduce framing bias.

Annotators were instructed to prioritize evidence sources referenced in the original fact-checking articles, as we consider these sources, curated by professional fact-checkers, to be more trustworthy. Nonetheless, annotators were also permitted to draw on evidence retrieved by our customized search engine. Using the list of common misinformative sources from [Schlichtkrull et al., 2023], we observed that 8 answers relied on a flagged source. However, this list was not applied during annotation, as it may be incomplete or contain false positives. Moreover, source credibility is often context-dependent, varying across topics and over time, which makes a static list insufficient for reliable filtering.

We acknowledge that some fact-checking articles included in our dataset may have been exposed to large pre-trained models during their pre-training phase. This is an open and ongoing challenge in constructing human-annotated datasets. To help mitigate this issue, the splits of our dataset are temporally organized. In particular, if the training data of a language model is cut off prior to the temporal start of our test set, then data leakage from pre-training into evaluation cannot occur.

We did not anonymize the data in AVerImaTeC, as all claims are derived from publicly available journalistic sources and primarily concern public figures and events. Preserving these references is essential to ensure the accuracy and integrity of fact-checking. Nevertheless, we recognize that some individuals may not wish to appear in the dataset. Accordingly, we have established an opt-out policy: if any individual featured in the dataset, as a claim speaker, person depicted in an image, subject of a claim, or author of a fact-checking article underlying a claim, wishes to be removed, they may submit a request, and the relevant content will be deleted from the dataset.

10 Conclusion

We present a real-world image-text claim verification dataset, annotated with QA pairs that capture the reasoning and evidence retrieval processes involved in claim verification. To ensure high annotation quality, we employed a multi-stage evidence sufficiency validation process, resulting in substantial inter-annotator agreement on both verdicts and QA annotations. In addition, we introduce a reference-based evaluation framework for open-web multimodal evidence retrieval, along with a set of baseline models for image-text claim verification that leverage web-sourced evidence. These contributions provide a foundation for advancing research in image-text claim verification.

Acknowledgement

This research was supported by the Alan Turing Institute and DSO National Laboratories in Singapore Partnership (ref DCfP2\100063). Zifeng Ding and Andreas Vlachos were further supported by the ERC grant AVeriTeC (GA 865958). Andreas Vlachos is also supported by the DARPA program SciFy. Michael Schlichtkrull is supported by the Engineering and Physical Sciences Research Council (grant number EP/Y009800/1), through funding from Responsible AI UK (KP0016).

References

Mubashara Akhtar, Michael Sejr Schlichtkrull, and Andreas Vlachos. Ev2r: Evaluating evidence retrieval in automated fact-checking. *CoRR*, abs/2411.05375, 2024.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification, FEVER@EMNLP*, pages 85–90, 2018.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *The Journal of Economic Perspectives*, 31(2):211–235, 2017. ISSN 08953309.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

Shivangi Aneja, Christoph Bregler, and Matthias Nießner. Catching out-of-context misinformation with self-supervised learning. *CoRR*, abs/2101.06278, 2021.

Sinan Aral and Dean Eckles. Protecting elections from social media manipulation. *Science*, 365(6456):858–861, 2019.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 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 4684–4696, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. New threats to society from free-speech social media platforms. *Commun. ACM*, 66(10):37–40, 2023.

Aviv Barnoy and Zvi Reich. The when, why, how and so-what of verifications. *Journalism Studies*, 20(16): 2312–2330, 2019.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Inf. Process. Manag.*, 56(5):1849–1864, 2019.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. DEFAME: dynamic evidence-based fact-checking with multimodal experts. *CoRR*, abs/2412.10510, 2024.

Rui Cao, Yuming Jiang, Michael Sejr Schlichtkrull, and Andreas Vlachos. Decompose and leverage preferences from expert models for improving trustworthiness of mllms. *CoRR*, abs/2411.13697, 2024.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 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), NAACL*, pages 3569–3587, 2024.

DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`.

Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7):pgae217, 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae217.

 https://doi.org/10.52202/085713-0154

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *CoRR*, abs/2405.11697, 2024.

Yi R. Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 1683–1698, 2021.

Max Glockner, Yufang Hou, and Iryna Gurevych. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5916–5936, 2022.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206, 2022.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL*, pages 493–503, 2019.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 893–903, 2023.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, and Ivan Nardini. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, pages 344–353, 2017.

Yoonsang Lee, Xi Ye, and Eunsol Choi. Ambigdocs: Reasoning across documents on different entities under the same name. *CoRR*, abs/2404.12447, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895, 2024.

Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020. doi: 10.1177/0022243719881113.

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

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. Open domain question answering with conflicting contexts. *CoRR*, abs/2410.12311, 2024a.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *CoRR*, abs/2406.08772, 2024b.

Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6801–6817, 2021.

Meta. Introducing llama 3.1: Our most capable models to date, 2024. URL `https://ai.meta.com/blog/meta-llama-3-1/`.

Shreyash Mishra, Suryavardan S, Amrit Bhaskar, Parul Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, and Asif Ekbal. FACTIFY: A multi-modal fact verification dataset. In *Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence AAAI*, volume 3199 of *CEUR Workshop Proceedings*, 2022.

Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*, pages 6149–6157, 2020.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558, 8 2021. Survey Track.

Eryn J. Newman, Maryanne Garry, Daniel M. Bernstein, Justin Kantner, and Stephen Lindsay. Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19:969–974, 2012.

Dan Saattrup Nielsen and Ryan McConville. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153, 2022.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 2532–2544, 2022.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. Qacheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 264–273, 2023a.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 6981–7004, 2023b.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multim. Inf. Retr.*, 13(1):4, 2024.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 5570–5579, 2025.

Karl Pearson and Olaus Magnus Friedrich Erdmann Henrici. 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, 1896.

Gordon Pennycook. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116:201806781, 2019. doi: 10.1073/pnas.1806781116.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. The art of SOCRATIC QUESTIONING: recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 4177–4199, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139, pages 8748–8763, 2021.

 https://doi.org/10.52202/085713-0154

Justus Randolph. Free-marginal multirater kappa (multirater free): An alternative to fleiss fixed-marginal multirater kappa. *Advances in Data Analysis and Classification*, 4, 01 2010.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 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, 2020.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

Yasmim Rocha, Gabriel Moura, Gabriel Alves Desiderio, Carlos Oliveira, Francisco Lourenço, and Larissa Nicolete. The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. *Journal of Public Health*, 31:1–10, 2021.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*, volume 36, pages 65128–65167, 2023.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6904–6913, 2023.

Jieun Shin and Kjerstin Thorson. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, 67(2):233–255, 2017. ISSN 0021-9916.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

Marek Strong, Rami Aly, and Andreas Vlachos. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 17021–17035, 2024.

S. Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Naresh Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. Factify 2: A multimodal fake news and satire news dataset. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI*, volume 3555 of *CEUR Workshop Proceedings*, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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, NAACL-HLT*, pages 809–819, 2018.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. "image, tell me your story!" predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 7845–7864, 2024.

Matthijs J. Warrens. Inequalities between multi-rater kappas. *Adv. Data Anal. Classif.*, 4(4):271–286, 2010. ISSN 1862-5347.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.

Qingzheng Xu, Heming Du, Huiqiang Chen, Bo Liu, and Xin Yu. MMOOC: A multimodal misinformation dataset for out-of-context news analysis. In *Information Security and Privacy - 29th Australasian Conference, ACISP*, volume 14897 of *Lecture Notes in Computer Science*, pages 444–459, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pages 2733–2743, 2023.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. Multimodal misinformation detection by learning from synthetic data with multimodal llms. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 10467–10484, 2024.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. 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 2099–2108, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the work are discussed in Section 8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

 https://doi.org/10.52202/085713-0154

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: There are no theoretical results in the paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiment settings and implementation details can be found in Section 7.1 (Model Implementation) and Appendix H.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Both the data and code are accessible to reviewers (see Appendix 5). As we anticipate using the dataset in a future shared task, we are as of submission time only releasing the training and development splits. We will make the test split available privately to reviewers upon request. We will make our data and code publicly available and maintain them on GitHub upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines: The information for data splits can be found in Section 5. Testing details are provided in Appendix H.1.

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

https://doi.org/10.52202/085713-0154

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our goal is not to advance in model design but to establish baselines for image-text claim verification. Besides, following the practice of [Schlichtkrull et al., 2023] and prior FEVER workshops (https://fever.ai/workshop.html), we do not require multiple runs of each model.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information for computation resources can be found in Appendix H.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential societal impact of the work has been discussed in Section 1 and Section 2.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

https://doi.org/10.52202/085713-0154

Answer: [Yes]

Justification: We used partially annotated data from AVeriTeC, AMMeBA and ClaimReview. They are explictly mentioned and discussed in Appendix A

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The dataset and code have been made available to reviewers. Our dataset and baseline will be made publicly available under a CC-BY-NC-4.0 license if accepted. Detailed dataset information can be found in Section 3, 4 and 5.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Data collection was carried out with the help of Appen (https://appen.com/). The payment details and annotator demographics can be found in Appendix C. Guidelines for annotators are provided Appendix B.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve human subjects. All annotators are recruited by a company, Appen (See Appendix C).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components. LLMs are only used for grammar error correction.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0154

A Statement of Dataset

We provide the training and development splits of our dataset for review.[3] However, due to naming inconsistencies with Kaggle, we are unable to upload the images there. Instead, we have made the image files available on Harvard Dataverse.[4] We cannot upload the entire dataset (including the .json file and images) to Harvard Dataverse, as it requires publishing the dataset when generating the Croissant file. To avoid unexpected inaccessibility of images on Harvard Dataverse, we also upload the whole development set (images and json file) to **supplementary**, due to file size limits.

Since we plan to use the test split for future shared tasks, we are withholding it at the time of submission. However, the test split will be made privately available to reviewers upon request.

If accepted, we will publicly release the training and development splits of the dataset and maintain both the data and baseline code on Harvard Dataverse and GitHub. The dataset and baselines will be licensed under the CC BY-NC 4.0 license.

B Annotation Platform and Guidelines

The annotation was performed using a custom-developed web platform specifically designed for this task from our team. We will make the platform's source code available upon request.

Annotators were provided with detailed guidelines outlining the annotation process. Due to the length of the guideline, we attached it to the **supplementary**. After reviewing the instructions, they underwent training for each phase. Specifically, we provide 14, 12 and 12 training instances for the first three phases, respectively. Phases four and five replicate phases two and three; therefore, no additional training instances were provided. We offered continuous feedback to annotators throughout both the training and annotation phases.

C Annotation Fees and Annotator Demographics

The annotation was conducted with the assistance of Appen,[5] a private company that provides machine learning services. Annotators were recruited through the company, which ensures fair compensation practices in accordance with their fair pay guidelines.[6] A total of 20 annotators participated in the project, 12 women (60%) and 8 men (40%). Fourteen annotators were based in the United States and six in the United Kingdom. Regarding age distribution, 20% of annotators were between 18 and 30 years old, 35% between 31 and 45, 25% were older than 45, and 20% chose not to disclose their age.

D Additional Dataset Information

In this section, we provide additional dataset statistics.

D.1 Statistics for Data Filtering in Phase One

We began with 2,353 fact-checking articles in Phase One (P1) of the annotation process. To increase the proportion of image-text claims, we incorporated partially annotated articles from three sources: 1) filtered articles from AVeriTec [Schlichtkrull et al., 2023] containing multimodal claims, 2) articles verifying image-related claims from AMMEBA [Dufour et al., 2024] and 3) all *true* claims from ClaimReview[7] over the past two years that included keywords such as *phtos* or *pictures*. Among the 2,353 articles, we identified 91 duplicate claims (e.g., those verified by multiple fact-checking organizations). Duplicates were detected automatically by computing cosine similarity and Jaccard distance between N-gram representations of textual part of claims. We manually reviewed claim pairs with a cosine similarity or Jaccard distance greater than 0.3. Despite these efforts, a small number of duplicates (fewer than 5%) remained, as fact-checkers may paraphrase similar claims differently.

[3] https://kaggle.com/datasets/a1ebcc27f233ca16e80fa7007d1a8cea6c491656d7b4ca6d87531df61c5089fd

[4] https://dataverse.harvard.edu/previewurl.xhtml?token=5fbd9e3c-bc9e-49d7-a109-eace6f1bdacd

[5] https://www.appen.com/

[6] https://success.appen.com/hc/en-us/articles/9557008940941-Guide-to-Fair-Pay

[7] https://developers.google.com/fact-check/tools/api

However, due to the temporal split of our dataset, such duplication does not cause data leakage across the splits. Paywalled or inaccessible articles were reported by annotators and subsequently discarded.

Although we leveraged coarsely annotated articles that verify multimodal claims, not all of them involved image-text relationships. In the first step, we filtered out claims of other modalities and retained only the image-text claims (around 85%) for further P1 annotation, which included image uploading, metadata collection, and verdict annotation. We further filter out image-text claims, where images are not used in claim verification (14.7%) or which involve image manipulations(4.0%). Both types were excluded from subsequent phases. Following [Schlichtkrull et al., 2023], we discard *speculative* claims, personal opinions and claims relying solely on *fact-checking reference* as the fact-checking strategy. After the P1 annotation, we retained a total of $\mathbf{1,457}$ valid image-text claims, for further processing in subsequent phases.

D.2 Discarded Claims

In Phase Two, annotators can report and skip invalid claims, particularly when images associated with claims are inaccessible due to regional restrictions. A total of 48 such cases were discarded.

As described in Section 4, when there are conflicting verdicts between annotators in Phases Two and Three, the affected claims proceed to Phases Four and Five for quality assurance updates and a second round of evidence sufficiency checks. In our dataset, 14.6% of claims required this second round of evaluation.

Claims with unresolved conflicts after the five-phase annotation process are also discarded. Overall, approximately 7% of claims were excluded due to irreconcilable disagreements in annotated verdicts, even after the second round of evidence sufficiency checks.

D.3 Metadata Distributions

Following [Schlichtkrull et al., 2023], we provide details of distributions of metadata of data splits in this part.

Table 5: **Distributions of claim types.**

Claim Type	Train	Dev	Test
Event/Property Claim	85.4	91.4	93.5
Causal Claim	5.7	0.7	3.4
Numerical Claim	3.2	3.9	1.4
Media Analysis Claim	21.1	14.5	30.1
Media Publishing Claim	2.1	2.0	2.0
Position Statement	1.2	0.7	0.3
Quote Verification	1.4	0.7	0.3

Table 6: **Distributions of fact-checking strategies.**

Claim Type	Train	Dev	Test
Media Source Discovery	26.4	20.4	26.4
Keyword Search	19.5	19.7	14.8
Written Evidence	87.5	89.5	84.7
Reverse Image Search	50.2	57.9	67.3
Image Analysis	20.7	24.3	21.3
Consultation	30.0	24.3	16.8
Video Analysis	6.8	4.6	8.2
Numerical Comparison	2.9	3.3	1.4
Satirical Source Identification	1.9	3.3	1.4
Fact-checker Reference	11.0	11.8	3.7
Geolocation	4.0	5.3	5.4

https://doi.org/10.52202/085713-0154

Metadata about Claims. We present the distribution of claim types and fact-checking strategies in Table 5 and 6, respectively. It is important to note that a single claim may belong to multiple claim types and can be verified using several fact-checking strategies. The most relevant locations

Table 7: **Counts of locations associated with claims.** Countries are represented with their ISO country code. We do not show countries with fewer than ten occurrences in the table, whereas the complete location information is available in our dataset.

Country code	Counts
IN	417
US	309
GB	74
UA	37
PK	30
IL	27
NG	20
LK	19
TR	18
PS	17
KR	17
KE	15
AU	14
RU	13
BD	12
JP	11
TH	10
MY	10
CN	10
CA	10
CR	10

associated with claims are listed in Table 7. We observed a bias in the geographic distribution of claims. As our dataset includes claims verified by fact-checkers, it may inherit any biases present in the original fact-checking sources [Shin and Thorson, 2017, Barnoy and Reich, 2019].

Table 8: **Distributions of question types.**

Question Type	Train	Dev	Test
Text-related	35.2	38.9	31.8
Image-related	63.4	58.8	62.2
Metadata-related	3.2	4.4	8.3
Commonsense-related	0.8	0.9	1.1

Metadata about QA Annotations. In addition to metadata for claims, we also provide metadata associated with the annotated QA pairs, including question types, answering methods, answer types and source mediums. Metadata statistics for questions and answers are presented in Table 8 and Table 9, respectively. Notably, a single question may belong to multiple question types.

E Inter-Annotator Agreement

To assess the quality of our annotated data, we recruited a different set of annotators to re-annotate 100 randomly sampled claims from our dataset and performed an inter-annotator agreement check. During re-annotation, we assumed that claim extraction and normalization had already been completed, and the annotators proceeded with the remaining phases. We ensured that the sample included at least five claims for each veracity label. The inter-annotator agreement check was done regarding both the agreement on verdicts and agreements on annotated QA pairs.

Table 9: **Distributions of metadata related to answers.** The first block is the distribution for answering methods, the second for answer types and the third for source medium.

Question Type	Train	Dev	Test
Image-search	54.6	54.6	51.9
Text-search	40.1	41.0	37.4
Metadata	1.8	3.0	7.7
Image Analysis	3.2	1.4	2.5
Abstractive	16.4	15.0	19.1
Extractive	57.9	61.3	58.9
Unanswerable	2.2	3.2	3.2
Boolean	21.9	18.8	17.3
Image	1.6	1.6	1.5
Web text	85.2	85.2	78.9
PDF	0.8	1.2	0.8
Metadata	1.2	2.1	7.1
Video	2.5	1.2	2.9
Image/graphic	4.4	5.3	4.6
Web table	0.3	0.5	0.1
Other	0.2	0.0	0.1

Verdict Agreement. To evaluate agreement on verdicts, we used Randolph's [Warrens, 2010] free-marginal multi-rater κ, which is well-suited for unbalanced datasets, following previous practices [Schlichtkrull et al., 2023, Ousidhoum et al., 2022]. We achieved an agreement score of 0.742 on the double-annotated claims.

QA Pair Agreement. For assessing agreement on annotated QA pairs, we recruited three best performing annotators to compare the similarities between the original and re-annotated QA annotations. Specifically, we instructed them to evaluate 1) whether the annotated verdict from our dataset could be supported by the original QA pairs; 2) how many original QA pairs for a claim are covered by the re-annotated QA pairs, and 3) how many QA pairs from the re-annotation are covered by the original QA pairs. The platform used for this agreement check is shown in Figure 3.

Justification Evaluation. Although we primarily focused on verdict and QA annotation agreement, we also conducted a small-scale human evaluation to assess the quality of justifications. We randomly sampled 20 justifications from our dataset and from [Schlichtkrull et al., 2023], and asked a human evaluator to rate them on a scale from 0 to 5. Our justifications achieved an average score of 4.2, compared to an average score of 1.95 from [Schlichtkrull et al., 2023].

F Human Alignment of Evaluation Metrics

Following [Schlichtkrull et al., 2023], we evaluate the quality of generated questions and retrieved evidence. Motivated by a recent study [Akhtar et al., 2024], we adopt a reference-based evaluation method that compares model responses to human-annotated ground-truth data. This method is applied to both question and evidence evaluations. To assess the reliability of the reference-based evaluation, we compare the resulting scores with human judgments obtained from independent raters (see Appendix F.1 and F.2). Additionally, we conduct *checklist* tests to evaluate the sensitivity of the reference-based evidence evaluation method, following the approach outlined in [Akhtar et al., 2024] (Appendix F.3).

F.1 Alignment Check on Question Evaluation.

Regarding question evaluation, we focus on assessing the semantics (i.e., textual content) of the questions. This evaluation approach aligns with that used in [Akhtar et al., 2024] and has also been

https://doi.org/10.52202/085713-0154

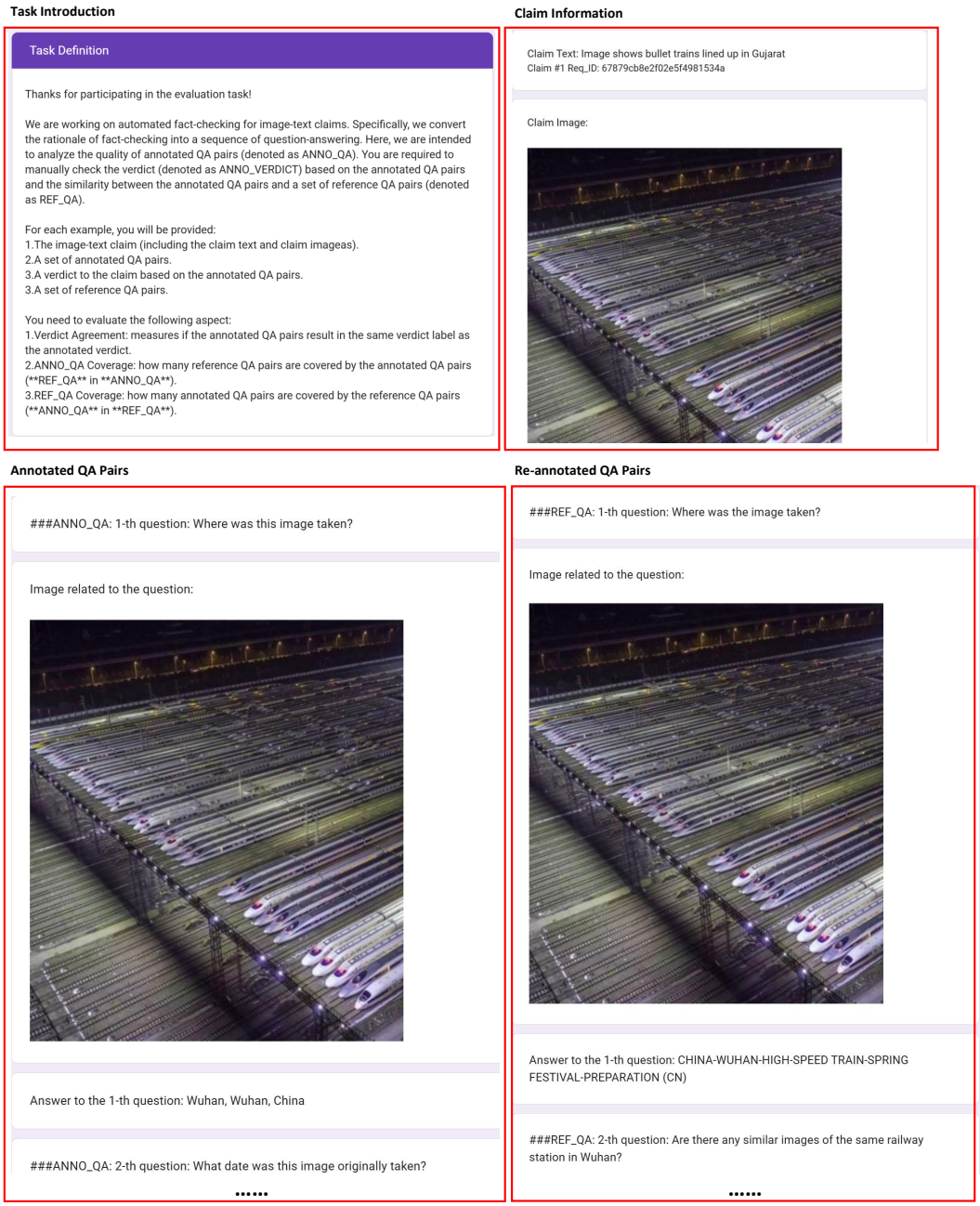

Figure 3: **Platform and instructions for validating annotators' agreement on QA annotations.**

applied at the FEVER workshop,[8] demonstrating its reliability and strong correlation with human ratings.

To further ensure alignment with human judgment in our setting, we conducted a small-scale human evaluation to validate the reference-based evaluation of questions. Specifically, we invited two NLP researchers (both authors of this paper) to participate in the assessment. They were presented with model-generated questions from the first 20 testing claims in our dataset and were asked to compare these questions against ground-truth annotated questions. The comparison was made based on two criteria: **Relevance** to claim verification (how relevant the generated questions are for verifying the claim) and **Coverage** of the ground-truth questions (how many ground-truth questions are accurately

[8]https://huggingface.co/spaces/fever/AVeriTeCFEVER8

covered by the predicted questions.). The agreement between human and the evaluation metric was measured using Spearman (ρ) [Spearman, 1904], achieving a value of **0.705**, and Pearson correlation coefficients (r) [Pearson and Henrici, 1896] achieving a value of **0.791**. These results indicate a strong alignment between human judgments and the reference-based evaluation scores on question evaluation.

Table 10: **Correlation between human evaluation and automatic evaluation metrics.** We presents both the Spearman (ρ) and Pearson (r) correlation coefficients.

Scorer	Text-only	Interleaved	Separate
ρ	0.263	0.08	0.332
r	0.286	0.14	0.381

F.2 Alignment Check on Evidence Evaluation.

We leveraged re-annotation and re-used the human judgment on QA annotations as described in Appendix E to validate the alignment between our automatic evidence evaluation method and human judgment. To assess alignment, we treated the original annotations from our dataset as predictions and the re-annotations as ground truth. We then applied the automatic evaluation methods to compare these *"predictions"* against the ground truth. Subsequently, we computed alignment scores by comparing the recall obtained from the automatic evaluation methods with human annotations. The correlation between the automatic evaluation scores and human assessments is presented using Spearman (ρ) and Pearson (r) correlation coefficients, as shown in Table 10. Our analysis shows that the separated evaluation method aligns most closely with human assessments, demonstrating the effectiveness of our approach.

Table 11: **Adversarial attack results for the assessment of the robustness of our evaluation method.** Results in the table are obtained by computing the evidence evaluation score difference (in %) between initial evidence and manipulated evidence with checklists.

Check.	Text-only	Interleaved	Separate
Completeness	-29.2	-19.8	-28.5
Shuffle	-10.2	4.1	-17.9
Irrelevant Image	0.0	9.4	-32.7
Inv. contract	-1.6	-7.4	1.5
Inv. num2text	1.8	21.3	-11.3
Inv. text2num	-0.3	24.5	0.6
Inv. synonyms	-11.8	88.8	-0.6
Redundant words	-3.1	85.4	-4.2
Fluency	-4.9	31.4	-0.7
Argument structure	-1.6	30.8	4.8
Image Invariance	0.0	20.0	-1.4

F.3 Robustness Check on Evidence Evaluation.

Motivate by [Akhtar et al., 2024, Ribeiro et al., 2020], we not only assess alignment but also validate the robustness of our evidence evaluation method using *checklist* tests. Notably, the sensitivity of reference-based evaluation for textual data has already been examined in [Akhtar et al., 2024], and we adopt their adversarial attack design on texts. Details for these attacks (*completeness*, *shuffle*, *Inv. contract*, *Inv. num2text*, *Inv. text2num*, *Inv. synonyms*, *Redundant words*, *Fluency*, *Argument structure*) can be found in the original study.

Given the multimodal nature of our evaluation, we extend the robustness tests to include visual adversarial checks. Specifically, we evaluate the robustness of the evidence evaluation method against two types of image-based perturbations: 1) **Irrelevant images**, replacing images in predicted evidence with unrelated images and 2) **Image invariance**, applying invariant manipulations, such as resizing and rotation, to images within the predicted evidence

https://doi.org/10.52202/085713-0154

We conducted robustness testing on 20 claims and compared the difference between the original evidence evaluation scores and the scores obtained after introducing adversarial evidence. The results are summarized in Table 11. A robust evaluation method is expected to show a significant performance drop when facing adversarial attacks in the first block of the table while maintaining consistent scores (i.e., minimal deviation from previous scores) when subjected to attacks in the second block.

The results indicate that both the Text-only and Separated reference-based evaluations exhibit robustness against textual adversarial attacks. In contrast, the interleaved evaluation is sensitive and unstable when faced with such attacks. This finding suggests that even advanced MLLMs, such as Gemini, may be prone to instability when handling interleaved image-text comprehension.

Regarding visual adversarial attacks (i.e., irrelevant image replacement and image invariance), the Text-only evaluation method fails to maintain stability, while our Separated evaluation method demonstrates robust performance.

G Additional Experimental Results

G.1 Zero-shot Performance of Baselines

Table 12: **Zero-shot performance of baselines. Q-Eval** and **Evid-Eval** denote for *recall* scores of generated questions and retrieved evidence, with reference of ground-truth questions and evidence. We report verdict prediction and justification generation scores conditioned on evidence retrieval performance, specifically only considering verdict accuracy and justification generation performance when the evidence score is above 0.2, 0.3 and 0.4.

LLM	**MLLM**	**Q-Eval**	**Evid-Eval**	**Veracity** (.2/.3/.4)			**Justifications** (.2/.3/.4)		
Paralleled Question Generation									
Gemini	Gemini	0.40	0.18	0.13	0.11	0.08	0.14	0.13	0.08
Qwen	Qwen-VL	0.41	0.16	0.08	0.07	0.04	0.11	0.09	0.07
Gemma	Gemma	0.37	0.22	0.12	0.12	0.08	0.17	0.15	0.10
Qwen	LLaVA	0.37	0.18	0.10	0.09	0.07	0.12	0.11	0.07
Dynamic Question Generation									
Gemini	Gemini	0.35	0.19	0.15	0.14	0.09	0.15	0.13	0.10
Qwen	Qwen-VL	0.30	0.14	0.11	0.10	0.06	0.11	0.09	0.06
Gemma	Gemma	0.27	0.19	0.12	0.11	0.07	0.14	0.13	0.08
Qwen	LLaVA	0.24	0.16	0.11	0.11	0.06	0.12	0.11	0.06
Hybrid Question Generation									
Gemini	Gemini	0.35	0.16	0.13	0.12	0.09	0.12	0.11	0.08
Qwen	Qwen-VL	0.34	0.14	0.09	0.08	0.05	0.10	0.09	0.05
Gemma	Gemma	0.26	0.25	0.15	0.14	0.09	0.19	0.18	0.12
Qwen	LLaVA	0.26	0.17	0.07	0.07	0.05	0.12	0.11	0.08

Due to the limitation of space, we provide the zero-shot performance of baselines in Table 12. The findings discussed in Section 7.2 hold for baselines under the zero-shot setting as well.

Table 13: **Justification evaluation scores with ROGUE-1 and Ev2R** when baselines are using ground-truth evidence.

LLM	**MLLM**	**ROGUE-1**	**Ev2R**
Gemini	Gemini	0.50	0.78
Qwen	Qwen-VL	0.44	0.68
Gemma	Gemma	0.43	0.75
Qwen	LLaVA	0.49	0.55

G.2 Justification Generation Scores with Ev2R

The justification evaluation scores of the baselines, even when provided with ground-truth evidence, are relatively low, as shown in Table 4. We attribute this to the limitations of ROUGE-1 in assessing open-ended generation. To address this, we adopted Ev2R [Akhtar et al., 2024], a reference-based evaluation method shown to perform well for open-ended generation tasks. The corresponding results are reported in Table 13.

A comparison between ROUGE-1 and Ev2R reveals that Ev2R produces significantly higher and more reasonable scores, which we find encouraging. As part of future work, we plan to conduct human alignment studies and incorporate such more sophisticated evaluation approaches for justification generation.

H Details of Experiments

H.1 Model Implementation

Hyper-parameters and Implementation. In the main experiments, we set the total number of generated questions to be 5 for all QG strategies. For HQG, the first two questions are generated in parallel while the rest three are generated exploited DQG. In cases of textual evidence retrieval (i.e., leveraging the tools of RIS and WST), we truncated retrieved texts into chunks with the maximum length of 128 and applied BM25 [Robertson and Zaragoza, 2009] to select the most relevant chunks to the given query. With the increasing capability of long-context understanding of existing LLMs, we keep the top 30 most related chunks, without a second stage of fine-grained re-ranking as what previous works have done [Chen et al., 2024, Schlichtkrull et al., 2023]. For the retrieved images returned by WSI, we compute their similarity scores with the given textual query with CLIP [Radford et al., 2021] and select the most related one as the image evidence source.

For the choice of LLMs in baselines, we have tried using LLaMA-3.1-8B-Instruct [Meta, 2024] as the LLM, whereas the model got stuck in loops, the same as reported by other users.[9] For LaVA-Next [Li et al., 2024], which is not designed for interleaved image-text, we only consider the textual part of evidence in verdict prediction and justification as we observed some issues with model generation with complex interleaved image-text information.

For the searching related tools, specifically WST, WSI and RIS, we used the API provided by Google. For web search with textual queries (WST and WSI), we first tokenize and post-tag words in queries and only keep verbs, nouns and adjectives as the search term [Karadzhov et al., 2017]. We set temporal constraints with input arguments, limiting all returned web pages published before claim dates. We keep the first 30 search results. For RIS, we employed the google cloud vision service for detecting web pages containing matched images with the querying image. However, the service does not embed arguments to set temporal constraints. Alternatively, we use a post-hoc method by leveraging the Python package *htmldate.fine_date* to filter out pages published before claim dates. We noticed a lot of web pages returned by RIS are social media posts, which are non-scrapable. For these pages, we use their page titles as the scraping content.

Few-shot Setting. To encourage models generate more critical questions for fact-checking, we exploit a computationally efficient method, few-shot learning. Specifically, we use a few training examples to guide models in question generation. Selecting similar claims to the inference one is important as similar claims may have similar reasoning path for claim verification. We rank the similarity between training claims and the inference claim with BM25 [Robertson and Zaragoza, 2009] by comparing their textual part. We set the number of shots to be 3 to balance between the input length and information from demonstrations.

For the PQG strategy, we directly provide models the ground-truth questions from selected demonstrations. For the DQG setting, models are provided the textual part of image-text claims and their first questions to generate the initial questions. For generating subsequent questions, each demonstration contains the textual part of a claim, its evidence history from previous QA pairs and the next question to be asked.

[9] https://www.reddit.com/r/LocalLLaMA/comments/1c858ac/llama3_seems_to_get_stuck_in_loops_sometimes/

https://doi.org/10.52202/085713-0154

Guidance for Tool Selection. Besides the guidance for tool selection in prompts as provided in Appendix I.4, we also used few demonstrations for inspire models to select proper tools. We provide few examples to guide tool selection as in the preliminary experiment we observed a heavy rely on VQA as the answering tool. We leveraged the metadata annotation of QA pairs, the answering method, for tool selection. *Image-search* will be mapped to RIS, *Text-search* with an image answer will be converted to WSI while with a textual answer will be mapped into WST. *Image analysis* will be converted to selecting VQA as the answering tool.

Computation Resources. All experiments are conducted with two GPUs each with 40G dedicated memory. Specifically, we exploited either A100 or L40 for our experiments. The Qwen + Qwen-VL baselines and the Qwen + LLaVA baselines take about three hours with A100 and Gemma-based baselines take about seven hours for inference on the test split. Models have a faster inference speed on L40, saving one third of inference time. The inference time of Gemini-based models varies, probably depending on the volume of API calls. Also, we observed instability of Gemini API (e.g., the API call returned *503*, saying the service is not available), maybe because of too many requests at the same time.

H.2 Experiment Environment and Packages

In this section, we introduce the experiment environment and packages in use. We implement all models under the PyTorch Library (version 2.4.0+cu121), with CUDA version 12.1. For the implementation of open-source LLMs and MLLMs, we leverage the HuggingFace Library, with the *Qwen/Qwen2.5-7B-Instruct* (Qwen), *google/gemma-3-12b-it* (Gemma), *Qwen/Qwen2.5-VL-7B-Instruct* (Qwen-VL) and *llava-hf/llava-v1.6-mistral-7b-hf* (LLaVA), respectively. The version of Huggingface is 4.50.2. For Gemini, we exploited the API, *gemini-2.0-flash-001*. For the CLIP model employed in image-text similarity computation for evidence rankin, we adopt the checkpoint, *openai/clip-vit-base-patch32*, from Huggingface as well.

I Prompts in Use

In this section, we provide the exact prompts in use for baselines.

I.1 Prompts for QA Conversion to Evidence Statement

Following [Akhtar et al., 2024], we convert QA pairs to evidence statement, for both evidence evaluation and maintaining the evidence history. We consider a text-only conversion for simplicity and use special tokens as placeholders for image. These placeholders could be placed with the exact images in the future. The prompt is demonstrated in Figure 4.

I.2 Prompts for Evaluation

We adopt a reference based evaluation strategy, which compare predictions against references, for both question evaluation and evidence retrieval evaluation.

Question Evaluation. Though questions could be multimodal, the semantics are the most informative. Therefore, we leverage a vanilla reference based evaluation scheme to compare the textual part of predicted questions and annotated questions. The exact prompt in use is shown in Figure 5.

Evidence Evaluation. We conducted a two-stage reference-based evaluation of evidence, as described in Section 6. In the first stage, we consider compare the textual of retrieved evidence and ground-truth evidence. Hence, the prompt used for evaluation is similar to that in question evaluation. The difference is that there are special image tokens in evidence and we need the evaluator to output the index of aligned predictions and ground-truth annotations. The prompt is illustrated in Figure 6.

I.3 Prompts for Question Generation

We considered three strategies for question generation as introduced in Section 7.1. The *hybrid* generation is the combination of the *paralleled* and the *dynamic* question generation strategy. Below, we provide prompts for the DQG and PQG strategies. For DQG, we use the prompt shown in Figure 7 and the PQG prompt is shown in Figure 8.

You are a expert writer. Given a question ([QUES]) and its answer [ANS], your goal is to convert the QA pair into a statement [STAT]. There could be images either in the question or the answer, which we use special tokens [IMG_1], [IMG_2] ... as placeholders for images. For instance, the question "When was the image published? [IMG]" asks for the publication date of the image denoted as [IMG]. Below are some examples:

[QUES]: What is the date of the claim?
[ANS]: Nov. 22, 2023.
[STAT]: The date of the claim is Nov. 22, 2023.

[QUES]: Did Trump pretended to be the palace guard in the meeting?
[ANS]: [IMG_1]
[STAT]: Trump dressed as [IMG_1] in the meeting.
……
[QUES]: When was the image shot? [IMG_1]
[ANS]: The image has been taken on Jan. 25, 1998.
[STAT]: [IMG_1] was taken on Jan. 25, 1998.

Please convert the QA pair below into its statement:
[QUES]: *[INFERENCE_QUES]*
[ANS]: *[INFERENCE_ANS]*
[STAT]:

Figure 4: **The prompt in use for converting QA pairs to evidence statement.**

You will get as input a reference question set ([REF]) and a predicted question set ([PRED]).Please verify the correctness of the predicted questions by comparing it to the reference questions, following these steps:1. Evaluate each question in the predicted question set individually: Check whether it is covered by any question in the reference set ([REF]). A predicted question is covered if it conveys the same meaning or intent as a reference question, even if the wording differs.2. Evaluate each question in the reference question set individually: Check whether it is covered by any question in the predicted set ([PRED]), using the same criteria. Do not use additional sources or background knowledge.3. Finally summarise (1.) Count how many predicted questions are covered by the reference questions and provide explanations([PRED in REF] and [PRED in REF Exp]), (2.) Count how many reference questions are covered by the predicted questions and provide explanations ([REF in PRED] and [REF in PRED Exp]).Generate the output as shown in the examples below:

[PRED]: 1. Is there a correlation between CO2 levels and climate change? 2. Where and when was the image taken? 3. What is the caption of the chart in the image? [REF]: 1. Was the source article publishing the chart discussing climate change? 2. Will a raise of CO2 levels lead to global warming? 3. Which country was shown in the image? 4. When was the image taken?[PRED in REF]: 2[PRED in REF Exp]: 1. The question is similar to the second reference question. 2. The question conveys similar information to the third and fourth question in the reference set. 3. The question is not covered by nor similar to any reference question.[REF in PRED]: 3[REF in PRED Exp]: 1. The question is not covered by the predicted question set. 2. The question is covered by the first predicted question. 3. The question is covered by the second question of the predicted questions. 4. The question is covered by the second predicted question.
……

Return the output in the exact format as specified in the examples, do not generate any additional output:
[PRED]: *[PRED_EVID]*
[REF]: *[REF_EVID]*

Figure 5: **The evaluation prompt for generated questions.**

For the few-shot question generation setting, we utilize the same prompts while adding a few demonstrations before the information of inference instances.

I.4 Prompts for Tool Selection

As mentioned in Section 7.2, we observed bias of models for heavily relying on VQA as the answering tool, diverging from fact-checkers' choice. This leads to failures for retrieving essential evidence.

https://doi.org/10.52202/085713-0154

You will get as input a reference evidence ([REF]) and a predicted evidence ([PRED]). [IMG_1], [IMG_2] .. are placeholders for images and they are regared as the same text token.Please verify the correctness of the predicted evidence by comparing it to the reference evidence. Note, a fact with "no answer could be found .." or "it is unkown .." contradicts with facts mentioning any exact information (i.e., indicating the answer can be found and it is known). Please verify following these steps:1. Evaluate each fact in the predicted evidence individually: is the fact supported by the REFERENCE evidence (reference evidence presents a similar)? Do not use additional sources or background knowledge.2. Evaluate each fact in the reference evidence individually: is the fact supported by the PREDICTED evidence? Do not use additional sources or background knowledge.3. Finally summarise (1.) how many predicted facts are supported by the reference evidence, which refernece evidence supports which predicted facts and explanations([PRED in REF] and [PRED in REF Exp]), (2.) how many reference facts are supported by the predicted evidence, which predicted evidence supports which reference fact and explanations ([REF in PRED] and [REF in PRED Exp]).Generate the output as shown in the examples below:

[PRED]: 1. The missle in [IMG_1] is Fateh 110. 2. Ilan Omar has attended the training in [IMG_2]. 3. Prince Phillip wore the Royal Guard uniform in Jan. 14, 2003. 4.The raid in Washington took place on Saturday, Oct. 26, 1999.[REF]: 1. [IMG_1] was taken in Jan. 20, 2003. 2. No evidence can be found related to the type of missle in [IMG_2]. 3. The woman in [IMG_3] for a training is not Ilan Omar. 4. No answer was found regarding when the raid in Washington took place. 5. Prince Phillip wore the Royal Guard uniform shown in [IMG_4] previously in Jan. 2003.[PRED in REF]: 1; (PRED_3,REF_5)[PRED in REF Exp]: 1. No relevant evidence to the fact can be found in the reference evidence set. 2. The fact contradicts with its relevant fact, the third fact, in the evidence set. It fact in the reference set claims the woman in the training is not Ilan Omar. 3. The fact is supported by the fifth fact in the evidence set. 4. The fact is refuted by the fourth fact in the reference set, which claims the date of the raid in Washington is unknown.[REF in PRED]: 2; (REF_1,RPED_3);(REF_5,PRED_3)[REF in PRED Exp]: 1. It is supported by the third fact in the predicted evidence. 2. It is refuted by the first fact in the predicted evidence set. 3. It is contracted with the second fact in the predicted evidence set, which claims the woman in the training is Ilan Omar. 4. It is refuted by the fourth fact in the predicted evidence which claims the date of the raid could be found. 5. The fact aligns with the third fact in the predicted evidence set.

……

Return the output in the exact format as specified in the examples, do not generate any additional output:
[PRED]: *[PRED_EVID]*
[REF]: *[REF_EVID]*

Figure 6: **The evaluation prompt for retrieved evidence.**

You are a fact-checker to ask questions to verify an image-text claim. Here is the image-text claim. The textual part is: *[CLAIM_TEXT];* with a list of images of the claim: *[CLAIM_IMAGE]*. We have already retrieved the evidence below: *[EVID_HISTORY].*
[IMG] is the placeholder for images in the evidence. Please ask another one question, either related to the textual part or related to the image part of the claim for the verification and avoid questions already presented in the evidence (though maybe no answer found to the question). For each question, please also indicate it is Text-related or Image-related before the question (using **Text-related:** and **Image-related:**). For image-related questions, please explicitly point out which image you are asking about (using **Image Index:**) and do not provide an index larger than the number of images (i.e., the index should be smaller than the total number of images). For instance, questions could be asked like:

Text-related: [QUES].\n**Image-related:** [QUES]. **Image Index:** 2.\n**Image-related:** [QUES]. **Image Index:** 2,3.\n[QUES] is the placeholder for the question to be generate.

The image index should be smaller than the total number of claim images. Please generate your question:

Figure 7: **The prompt for dynamic question generation.** In the few-shot setting, the second paragraph is replaced with the few-shot demonstrations.

Considering the issue, besides the tool definitions, we provide a few demonstrations, each consisting of a question, a question type and the tool should be selected. Specifically, we leverage the annotated metadata information of questions. For questions annotated with the *answering method* of image-search, we consider the tool of RIS for such cases. For questions with the answering method as

You are a fact-checker to ask questions to verify an image-text claim. . Here is the image-text claim. The textual part is: *[CLAIM_TEXT];* with a list of images of the claim: *[CLAIM_IMAGE]*.

Please ask *[NUM_QUES]* questions related to the textual part or the image part of the claim for the verification of the claim. For each question, please also indicate it is Text-related or Image-related before the question. For Image-related questions, please also indicate which images the question is about with **Image Index:**. Specifically, if the question is related to the first and second image of the claim, it should be **Image Index:** 1,2.

We illustrate an example of first three questions as below ([QUES] is the placeholder for the question.):
1..**Text-related:** [QUES]
2. **Image-related:** [QUES] **Image Index:** 1.
3. **Image-related:** [QUES] **Image Index:** 2.

Please generate *[NUM_QUES]* questions:

Figure 8: **The prompt for paralleled question generation.** In the few-shot setting, the third paragraph is replaced with the few-shot demonstrations.

You need to select a proper tool to answer a given question. We have four tools:
(A). Reverse image search: The tool aims to find information related to given images, such as the date of, the event in or the celebrities in the image. It uses the provided images as input to query a search engine.
(B). Visual question answering: The tool is frequently used for answering questions related to image details (e.g., the text in the image and the weather in the image).
(C). Searching Google using texts for textual information: The tool aims to find related textual information by querying a search engine with texts.
(D). Searching Google using texts for images: The tool aims to find related images by querying a search engine with texts.
You need to response with the options for tools (i.e., A, B, C or D) and please do not respond with any other words. For text-related questions, please select from tool C and D. For image-related questions, please select from A and B.
Below we provide some examples:

[DEMONSTRATIONS]

Question: *[INFERENCE_QUES]* Question type: *[INFERENCE_QUES_TYPE]* Tool option:

Figure 9: **The prompt in use for selecting tools to answer questions.** [DEMONSTRATIONS] are placeholders of examples provided to guide the tool selection.

text-search while the answers are not images, we regard WST as the tool to be selected; if there are image answers, then the WSI should be the tool. For questions answered by image analysis, we would consider VQA as the answering method.

The prompt for tool selection is shown in Figure 9.

I.5 Prompts for Answer Generation

As introduced in Section 7.1, when leveraging the tools of RIS, WST and WSI, there follows an answering model (either an LLM or an MLLM) to leverage retrieved evidence to address the question.

For using an LLM to leverage textual evidence, we use the prompt: *You need to answer a question according to a set of retrieved documents. Question:* `[QUES]`; *Document:* `[RETRIEVED_DOC]`. *If the question is not answerable according to the provided document, please answer as: No answer can be found. Start you answer as: **ANSWER:***

https://doi.org/10.52202/085713-0154

For VQA with an MLLM, we prompt models with the template below for an answer: *Question:* `[QUES]`; *Related images to the question:* `[IMAGES]`.

I.6 Prompts for Verdict Prediction

The verifier receives the claim and the retrieved evidence for predicting a veracity label of the claim. Below is the prompt exploited for the verifier: *You need to select a verdict for a given Image-Text claim when provided a set of evidence. [IMG] is a placeholder for images. We provide four verdict labels and the definitions of them below: Supported: The claim is supported by the evidence presented. Refuted: The claim (either the text or the image part) is contradicted by the evidence presented. Not Enough Evidence: There is not enough evidence (NEE) to support or refute the claim. Conflicting: The claim is misleading due to conflicting evidence/cherry-picking, but not explicitly refuted. You need to response with the verdict for the claim (i.e., Supported, Refuted, Not Enough Evidence or Conflicting) and please do not respond with any other words. The metadata of the claim:* `[DATE_AND_LOCATION]`. *Claim:* `[CLAIM_TEXT]`; *Claim images:* `[CLAIM_IMAGES]`. *Here is the evidence:* `[EVID]`. *Verdict:*

I.7 Prompts for Justification Generation

The prompt for justification generation receives the information about the claim (textual part and claim images), retrieved evidence and predicted verdict to explain how the verdict could be reached. Below is the exact prompt in use:

Given an image-text claim and a set of evidence for verifying the claim, a fact-checker predict a veracity label for the claim. You need to explain how the verdict is reached for the image-text claim. Below is information for the image-text claim: The metadata of the claim: `[DATE_AND_LOCATION]`. *Claim:* `[CLAIM_TEXT]`; *Claim images:* `[CLAIM_IMAGES]`. *The predicted verdict is:* `[PRED_VERDICT]`. *Here is the evidence:* `[EVID]`. *Please generate your justification (i.e., explanation) for the verdict:*

Outputs from MLLMs are verbose, whereas human annotated justifications are concise. Therefore, we conduct one step further to prompt the corresponding LLMs to summarize the generated jsutifications in one or two sentences.