M4FC: a Multimodal, Multilingual, Multicultural, Multitask Real-World Fact-Checking Dataset

Jiahui Geng[*1], **Jonathan Tonglet***[2,3,4], **Iryna Gurevych**[1,2]
[1] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science, TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE, Germany
[3] Department of Electrical Engineering, KU Leuven, Belgium
[4] Department of Computer Science, KU Leuven, Belgium
www.ukp.tu-darmstadt.de

Abstract

Existing real-world datasets for multimodal automated fact-checking have multiple limitations: they contain few instances, focus on only one or two languages and tasks, suffer from evidence leakage, or depend on external sets of news articles for sourcing true claims. To address these shortcomings, we introduce M4FC, a new real-world dataset comprising 4,982 images paired with 6,980 claims. The images, verified by professional fact-checkers from 22 organizations, represent diverse cultural and geographic contexts. Each claim is available in one or two out of ten languages. M4FC spans six multimodal fact-checking tasks: visual claim extraction, claimant intent prediction, fake detection, image contextualization, location verification, and verdict prediction. We provide baseline results for all tasks and analyze how combining intermediate tasks influence downstream verdict prediction performance. We make our dataset and code available.[1]

1 Introduction

Since 2020, approximately 80% of online misinformation verified by fact-checkers has been multimodal, combining text with images, videos, or audio (Dufour et al., 2024). Debunking multimodal misinformation is challenging and time-consuming for human fact-checkers (Silverman, 2013; Mossou and Higgins, 2021; Khan et al., 2023, 2024). It typically involves performing several sub-tasks and utilizing specialized tools like satellite imagery providers and sun angle calculators (Mossou and Higgins, 2021; Khan et al., 2023, 2024). Additionally, it requires expert knowledge of specific languages and cultural contexts (Silverman, 2013). This work specifically targets the two most common categories of image-based misinformation (Dufour et al., 2024): (1) out-of-context (OOC) misinformation, where authentic images are paired with false claims about their date, location, or the people and events depicted; and (2) false claims paired with manipulated or fake images.

Multimodal automated fact-checking (AFC) assists human fact-checkers by automating parts of their workflow (Akhtar et al., 2023). Several datasets have been proposed for multimodal AFC. Synthetic datasets, on the one hand, are large and scalable (Luo et al., 2021; Shao et al., 2024) but cannot fully mimic real-world misinformation. On the other hand, real-world datasets consist of claims verified by human fact-checkers, but these datasets tend to be much smaller (Zlatkova et al., 2019; Hu et al., 2023; Papadopoulos et al., 2024; Tonglet et al., 2024). Existing real-world datasets suffer from additional limitations. While multimodal misinformation is a global problem, these datasets cover only one or two languages and lack geographic diversity. Furthermore, the human fact-checking workflow involves several intermediate tasks before providing a verdict. However, the real-world datasets fail to account for most of these intermediate tasks.

In this work, we introduce **M4FC**, a **m**ultimodal, **m**ultilingual, **m**ulticultural, and **m**ultitask real-world **f**act-**c**hecking dataset, with 4,982 images and 6,980 claims. M4FC addresses several limitations of existing resources. It contains labels for six multimodal AFC tasks organized into a pipeline, as illustrated in Figure 1. This pipeline allows for automating a larger part of the fact-checking workflow and studying the interaction between intermediate tasks and verdict prediction. Images, claims, task labels, and metadata are drawn from articles produced by 22 fact-checking organizations across 17 countries. As a result, M4FC offers broad cultural and geographic coverage and spans 10 claim languages: Arabic, Dutch, English, French, German, Portuguese, Spanish, Tamil, Telugu, and Turkish. We introduce two new multimodal AFC tasks.

*These authors contributed equally to this work.
[1]github.com/UKPLab/M4FC

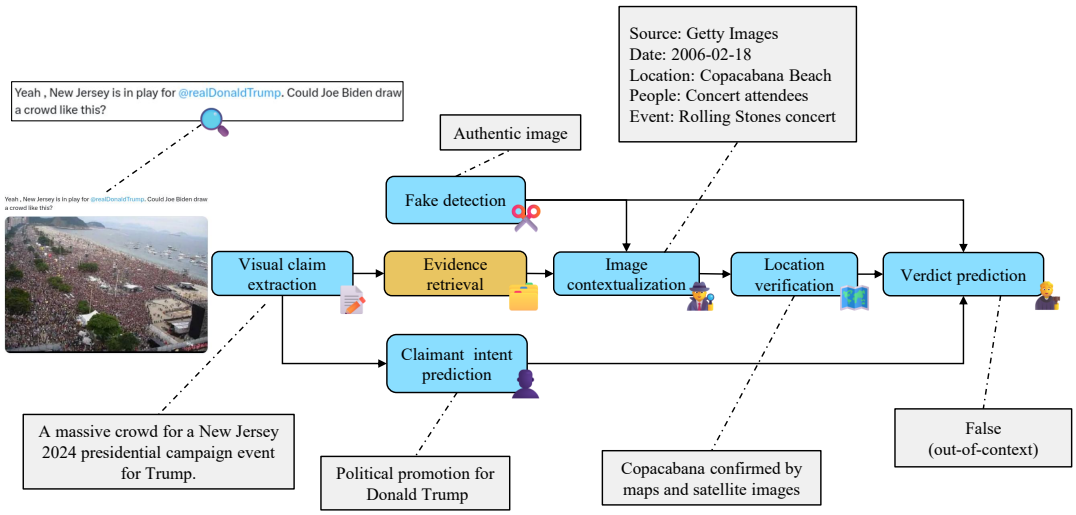


Figure 1: An out-of-context instance from M4FC. The six AFC tasks are shown in blue and their output in gray.

Visual claim extraction addresses the mandatory step of formulating verifiable claims based on a screenshot of a social media post containing the image. *Location verification* validates candidate locations for images using two types of evidence previously overlooked in multimodal AFC research: maps and satellite images. We provide baseline results for all tasks. Furthermore, we evaluate how different sequential combinations of intermediate tasks affect verdict prediction performance.

In summary, our contributions are as follows: (1) we introduce M4FC, a large-scale real-world multimodal, multilingual, multicultural, and multitask fact-checking dataset; (2) we introduce two new challenging multimodal AFC tasks; and (3) we conduct comprehensive experiments, including analyses of how combining intermediate tasks impacts verdict prediction performance.

2 Related work

2.1 Synthetic multimodal AFC datasets

Several synthetic datasets have been proposed for multimodal AFC (Sabir et al., 2018; Müller-Budack et al., 2020; Luo et al., 2021; Shao et al., 2024; Liu et al., 2025). While synthetic datasets are large and scalable, they often suffer from biases. Papadopoulos et al. (2025) demonstrated that simple models can exploit these biases to achieve near SOTA performance, raising concerns about the reliability of synthetic data for evaluating progress in multimodal AFC.

2.2 Real-world multimodal AFC datasets

Real-world datasets, summarized in Table 1, collect claims from fact-checking articles. Because human fact-checking is time-intensive, limited articles are available, and real-world datasets are smaller than synthetic ones. They also often lack true claims (Nielsen and McConville, 2022), as fact-checkers prioritize articles about false claims. To compensate, many datasets sample image-caption pairs from reputable news outlets and treat them as true claims (Jin et al., 2017; Zlatkova et al., 2019; Aneja et al., 2023; Hu et al., 2023). However, this introduces a distribution mismatch (Nielsen and McConville, 2022), as the content, resolution, and topics of news images may differ from those in online misinformation. Existing datasets cover one or two claim languages and provide labels for one task: verdict prediction. Exceptions include 5Pils (Tonglet et al., 2024), which addresses the task of image contextualization, and 5Pils-OOC (Tonglet et al., 2025), which covers both tasks. Many datasets gather evidence via reverse image search (RIS) (Zlatkova et al., 2019; Tonglet et al., 2024) or by querying a search engine (Hu et al., 2023; Papadopoulos et al., 2024; Cao et al., 2025). To ensure realistic evaluation, evidence written by fact-checking organizations or published after the claim date should be excluded, a filtering step applied by some datasets (Zlatkova et al., 2019; Tonglet et al., 2024; Cao et al., 2025). However, this cannot prevent the risk that leaked evidence are part of the

Dataset	# tasks	# images	Claims				Evidence	
			# claims	# languages	Data split	Source	Leak?	Type
Weibo (Jin et al., 2017)	1	9,528	9,528	1 (Zh)	Event	FC & News	-	-
Fauxtography (Zlatkova et al., 2019)	1	1,233	1,233	1 (En)	Random	FC & News	No	Text
COSMOS (test) (Aneja et al., 2023)	1	1,700	3,400	1 (En)	No split	FC & News	-	-
MR2 (Hu et al., 2023)	1	14,700	14,700	2 (En, Zh)	Event	FC & News	Yes	Text, Image
Ookpik (Pham et al., 2024)	1	545	1,090	1 (En)	No split	FC & News	-	-
Post-4V (Geng et al., 2024)	1	186	186	1 (En)	No split	FC	-	-
VERITE (Papadopoulos et al., 2024)	1	662	1,000	1 (En)	No split	FC	Yes	Text, Image
5Pils (Tonglet et al., 2024)	1	1,676	1,676	1 (En)	Temporal	FC	No	Text
5Pils-OOC (Tonglet et al., 2025)	2	624	1,248	1 (En)	No split	FC	No	Text, Image
AVerImaTeC (Cao et al., 2025)	1	1,297	1,297	1 (En)	Temporal	FC	No	Text, Image
XFacta (Xiao et al., 2025)	1	2,400	2,400	1 (En)	Random	CN & News	Unknown	Text, Image
M4FC (ours)	**6**	**4,982**	**6,980**	**10**	**Temporal**	**FC**	**No**	**Text, Map, Satellite**

Table 1: Comparison of real-world multimodal AFC datasets. En and Zh indicate claims in English and Chinese, while FC, News, and CN stand for fact-checking articles, news articles, and X Community Notes, respectively.

pre-training corpus of MLLMs (Xiao et al., 2025).

M4FC addresses several limitations of existing datasets. By gathering data from 22 fact-checking organizations from around the world, it covers 10 languages and is geographically diverse. Furthermore, M4FC includes six tasks, allowing for the combination of intermediate tasks with verdict prediction and aligning multimodal AFC research more closely with human fact-checking practices.

3 M4FC Dataset

3.1 Six multimodal AFC tasks

M4FC includes six tasks, including two new ones, visual claim extraction and location verification.

Multimodal claims are not always expressed online in an explicit and verifiable format. **Visual claim extraction (VCE)** addresses this issue by generating an English-language claim from the screenshot of a social media post. This new task involves two challenges: (1) extracting visual cues and text from a wide range of languages, and (2) leveraging external knowledge. As shown in Figure 2, parts of the claim appear in the text, others on the building within the image, and further clarification requires external knowledge.

Claimant intent prediction (CIP) seeks to identify the underlying purpose of the claimant, such as expressing sarcasm or promoting a political agenda. We formulate it as a free-text generation task. While Da et al. (2021) and Wu et al. (2025) predicted claimant intents using synthetic data, we extend their work to real-world claims.

Fake detection (FD) identifies images that are manipulated or fake. Manipulations include face swaps or background changes, while fake images encompass AI-generated content and human-made digital artwork. This classification task has two labels: *authentic* and *manipulated/fake*.

Image contextualization (IC) determines the original context of an image (Tonglet et al., 2024). This task is framed as predicting a set of items. We cover the following context items (Tonglet et al., 2024, 2025): *provenance* (binary variable indicating whether the image has previously appeared on the web), *source* (original author or publisher), *date*, *location*, *motivation* (why the source made or published the image), *people*, *things*, and *event*.

Human fact-checkers often compare the image with aerial views of a candidate location to further validate it (Khan et al., 2024). Inspired by this practice, we define the task of **location verification (LV)**: given a map or satellite view of a candidate location, the goal is to determine whether it is consistent with the target image. In the example of Figure 3, the first candidate location, Sart Canal Bridge, is confirmed by the matching road network below the bridge. In contrast, another candidate, the Ringvaart aqueduct, is ruled out because it spans a four-lane highway, which does not match the content of the image. This task relates to cross-view geolocalization (Lin et al., 2013), where the objective is to retrieve a matching satellite image from a large database, given a ground-level image.

Verdict prediction (VP) classifies a claim as true or false. The task is defined in four settings, based on two factors: whether claims are *English-only* or *multilingual*, and whether the verdicts are *imbalanced* or *balanced*. In the *English-only* setting, all claims are in English, while the *multilingual* setting covers 10 languages. The *imbalanced* setting captures the real-world distribution of fact-checking verdicts, in which the majority of claims are false. The *balanced* setting is constructed by replacing a portion of false claims with synthetic

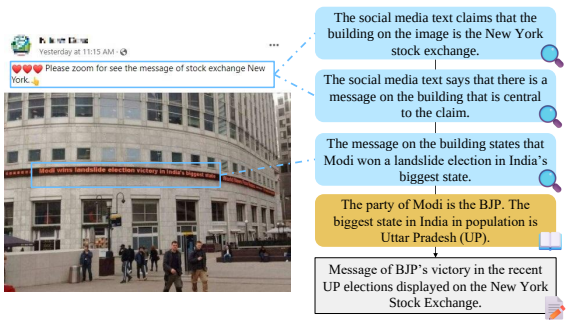


Figure 2: Illustration of the visual claim extraction task. Blue, yellow, and gray boxes indicate visual cues, external knowledge, and the extracted claim, respectively.

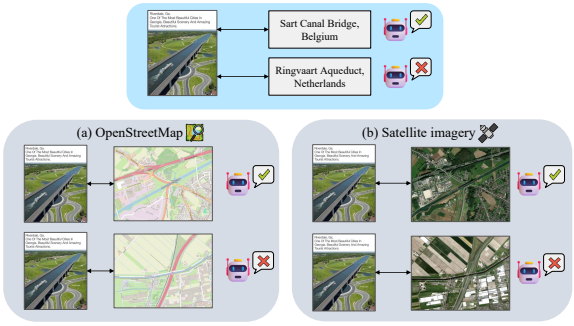


Figure 3: Illustration of the location verification task. Satellite images are sourced from ESRI World Imagery.

true claims linked to the same image.

3.2 Dataset construction

3.2.1 Data collection and labeling

We collected articles from 22 organizations, based in 17 different countries and all members of the International Fact-Checking Network (IFCN)[2], ensuring high geographic and linguistic diversity and adherence to professional standards. Table 2 reports the distribution of the images per fact-checking organization. Some organizations produce a higher volume of articles than others.

Following Tonglet et al. (2024), we use GPT4o (OpenAI, 2023) to extract claims in English and in the original language, task labels, and metadata from the article. GPT4o also generates synthetic true claims based on the context items of images labeled as false (out-of-context). For a random half of those instances, the original false claim is replaced by the synthetic true claim under the *balanced* setting. For the *multilingual* setting, we consider languages with at least 40 claims; claims in less-represented original languages are translated in English. All prompts are provided in Appendix A. Verdicts are mapped to the following categories: true, false, mixture, unverified. Finally, we apply a temporal split to the dataset to avoid temporal leakage (Glockner et al., 2022). Appendix B provides more details about the data collection and labeling.

3.2.2 Data validation

We adopt the validation procedure from Tonglet et al. (2024) to assess the quality of GPT4o annotations. For each organization, we randomly sample up to 20 articles. Each article is reviewed by two annotators who evaluate the correctness of the data fields extracted by GPT4o. Annotators are native speakers of the organization's language and are recruited through Prolific.[3] 79 annotators participated in the study, of which 33 were excluded for failing more than five out of 15 attention checks.

During post-processing of the Prolific results, we distinguish between incorrect and missing labels. A label is marked as missing when GPT4o outputs "not enough information", despite relevant information being present in the article. A label is considered incorrect or missing only if both annotators agree on it. Under this criterion, the majority of labels are accurate, with average correctness per field ranging from 86.2% (for claimed date) to 99.31% (for verdict). Using a more lenient criterion, counting a label as incorrect if at least one annotator flags it, the correctness ranges from 49.6% (for claimed date) to 89.63% (for verdict), still indicating strong automated annotation quality.

We compute inter-annotator agreement (IAA) using Randolph's κ (Randolph, 2005), which ranges from 0.32 to 0.81 across all data fields, indicating moderate to strong agreement. If we exclude instances labeled by GPT4o as not "enough information", the main source of disagreement, the IAA ranges from 0.45 to 0.85. Guidelines and detailed scores are provided in Appendices C.1 and C.2.

3.2.3 Dataset statistics

Table 3 provides the distribution of instances and the type of inputs per task. 1,148 images are unsuitable for visual claim extraction because they do not contain any embedded text, making it impossible to infer a claim. Location verification has fewer instances than other tasks, and may benefit from dedicated synthetic datasets in future work. Additional statistics are provided in Appendix D.

[2] ifcncodeofprinciples.poynter.org/signatories

[3] prolific.com

Organization	Country	Language	# articles
factly.in	India	English, Kannada, Telugu	1027 (20.6%)
fatabyyano.net	Jordan	Arabic	658 (13.2%)
pesacheck.org	Kenya	English, French	561 (11.3%)
snopes.com	USA	English	454 (9.1%)
indiatoday.in	India	English	333 (6.7%)
leadstories.com	USA	English	328 (6.6%)
poligrafo.sapo.pt	Portugal	Portuguese	284 (5.7)%)
congocheck.net	DRC	French	260 (5.2%)
larepublica.pe	Peru	Spanish	253 (5.1%)
correctiv.org	Germany	German	220 (4.4%)
dfrac.org	India	English	104 (2.1%)
knack.be	Belgium	Dutch	95 (1.9%)
mythdetector.ge	Georgia	English	67 (1.3%)
dubawa.org	Nigeria	English	65 (1.3%)
youturn.in	India	Tamil	59 (1.2%)
dogrula.org	Turkey	Turkish	52 (1.0%)
211check.org	South Sudan	English	51 (1.0%)
verify-sy.com	Turkey	Arabic	42 (0.8%)
tjekdet.dk	Denmark	Danish	29 (0.6%)
factcheckcenter.jp	Japan	Japanese	18 (0.4%)
factchecker.gr	Greece	Greek	17 (0.3%)
factchecklab.org	Hong Kong	Chinese	5 (0.1%)

Table 2: M4FC data source distribution.

3.2.4 Evidence collection

For image contextualization and verdict prediction, we collect up to 20 web evidence using the Google RIS engine.[4] For location verification, we collect maps from OpenStreetMap (OpenStreetMap contributors, 2017) and satellite images from ESRI World Imagery, [5] accessed with staticmap.[6]

4 Experiments

4.1 Evaluation metrics

For visual claim extraction and claimant intent prediction, we assess the generated texts using Rouge-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). For fake detection, location verification, and verdict prediction, we report the macro F1-score (F1). For image contextualization, following Tonglet et al. (2025), we adopt METEOR for source, motivation, things, and event, F1 for people, Δ for date, and $CO\Delta$ for location. Δ and $CO\Delta$ measure the inverse proportional distance from the ground-truth in years and in thousand kilometers, respectively.

4.2 Baselines

For fake detection, we evaluate three specialized detectors, Deep-Fake Detector (Prithiv, 2024), AI-Image Detector (Dafilab, 2025), and RealFake Detector (Mikedata, 2025), considering both their pre-trained versions and fine-tuned variants on the MF4C training set. For other tasks, we employ three widely adopted open-weight MLLMs, InternVL2.5-8B (Chen et al., 2024), Qwen2.5-VL-7B (Qwen-Team, 2025), and Llama-3.2-11B-Vision (MetaAI, 2024), alongside two proprietary models, GPT4o-mini (OpenAI, 2023) and Gemini-1.5-Flash (Gemini-Team, 2024), with task-specific prompts provided in Appendix E. In addition, for image contextualization and verdict prediction, we benchmark against the current state-of-the-art (SOTA) methods, COVE (Tonglet et al., 2025) and DEFAME (Braun et al., 2025), respectively. COVE combines two models: LlavaNext-7B (Liu et al., 2023, 2024) and Llama-3-8B (MetaAI, 2024). For DEFAME, we use GPT4o-mini as the backbone LLM. Appendix F provides implementation details.

[4] cloud.google.com/vision/docs/detecting-web
[5] arcgis.com/home/item.html
[6] github.com/komoot/staticmap

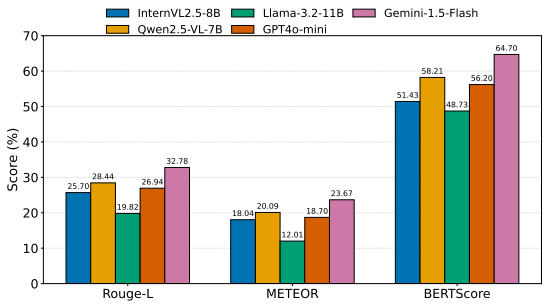
Task	# instances	# train	# dev	# test	Input
VCE	3,839	2,285	415	1,144	I
CIP	1,789	1,113	164	512	I, C
FD	4,954	2,978	498	1,478	I
IC	4,720	2,887	480	1,353	I, C, E
LV	195	96	27	72	I, E
VP	4,955	2,979	498	1,478	I, C, E

Table 3: Number of instances and input types per task. I is the image, C is the claim, E is the evidence set.

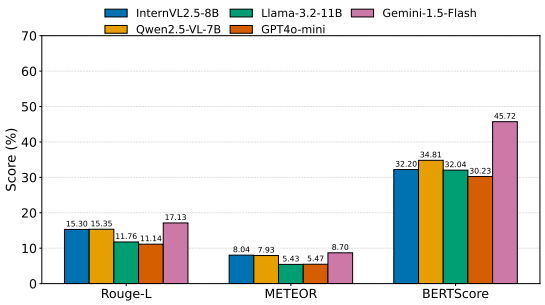
4.3 Baselines results

We analyze results for all tasks on the test set. Additional result tables are provided in Appendix G.

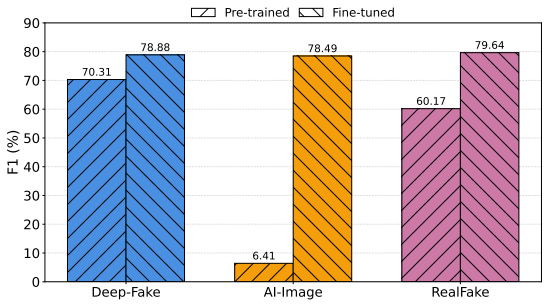
Visual claim extraction Figure 4a shows that most models achieve decent performance, with the exception of Llama-3.2-11B-Vision, while Gemini-1.5-Flash consistently outperforms all others. Rouge-L and METEOR scores suggest a strong alignment between generated and reference claims, whereas BERTScore indicates a more moderate level of semantic correspondence. We manually analyzed 50 randomly sampled predictions from Gemini-1.5-Flash. Among them, 38 claims are accurate, nine were partially correct but lacked external world knowledge, and three were incorrect due to the model's inability to capture sarcasm in social media posts. Figure 5 presents two error cases: on the left, the model fails to connect the image with its broader context, although the claim publication date and the landscape hint at a connection with the Maui fires; on the right, the text in German translates to "*There are over 100 years between the two photos. You can clearly see the ever-rising sea level, can't you?*", yet the model overlooks the sarcasm and extracts a claim that directly contradicts the ground truth.



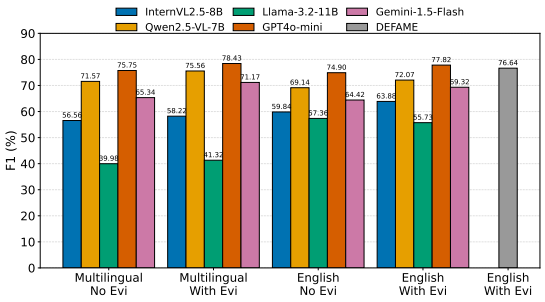
(a) Visual claim extraction.



(b) Claimant intent prediction.



(c) Fake detection.



(d) Verdict prediction.

Figure 4: Visual claim extraction (VCE), claimant intent prediction (CIP), fake detection (FD), and verdict prediction (VP) results (%). All VP results are shown in the *balanced* setting.

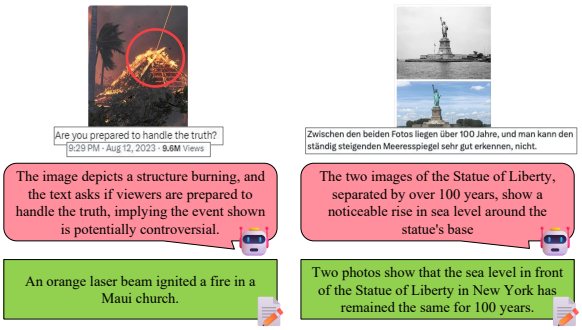


Figure 5: Visual claim extraction error examples with Gemini-1.5-Flash on M4FC test.

Claimant intent prediction Figure 4b shows that Gemini-1.5-Flash achieves the best performance, with a clear margin on BERTScore compared to other models. Llama-3.2-11B trails slightly behind. GPT4o-mini underperforms because it often refuses to execute the task, flagging claimant intent prediction as a potential policy violation. The overall scores are low, highlighting that real-world claimant intent prediction remains challenging for current MLLMs.

Fake detection Figure 4c reports F1 scores for specialized fake detectors. While the pre-trained Deep-Fake detector achieves a relatively strong F1 of 70.31%, AI-Image detector is almost ineffective at only 6.41%, and RealFake detector remains modest at 60.17%. Although the M4FC train test is small and does not overlap in time with the test set, fine-tuning on it yields high performance gains across all models. This means that features can be learned from prior real-world cases of manipulated and fake images and generalized to newer cases.

Image contextualization As shown in Table 4, performance varies substantially across different context items, with no model dominating all categories. Qwen2.5-VL-7B achieves the highest scores on Date (30.3%), Location (45.0%), Motivation (16.8%), and Event (16.2%). In contrast, Gemini-1.5-Flash demonstrates strength in recognizing entities, leading in People (45.0%) and Things (22.7%). The large gaps observed in Location and People further reveal that these dimensions are particularly challenging for current MLLMs. COVE does not provide better results than most MLLMs. On the contrary, it is outperformed by all MLLMs for Source and Date. All models struggle with Source. Consistent with prior work, predicting the date of the image is more challenging than predicting its location (Tonglet et al., 2024, 2025).

Location verification We evaluate MLLMs by comparing each claim image with candidate loca-

	Source (M)	Date (Δ)	Loc. (CO Δ)	Mot. (M)	People (F1)	Things (M)	Event (M)
InternVL-2.5-8B	4.8	21.6	26.5	9.5	24.4	13.4	8.0
Qwen2.5-VL-7B	5.2	**30.3**	**45.0**	**16.8**	41.0	13.7	**16.2**
Llama-3.2-11B	**7.7**	29.1	38.6	10.5	34.6	14.0	12.2
GPT4o-mini	5.7	20.4	28.9	7.6	20.4	13.3	8.3
Gemini-1.5-Flash	7.1	23.9	42.3	16.6	**45.0**	**22.7**	15.8
COVE (Llama-3-8B)	1.7	17.3	33.0	16.2	34.5	12.1	14.4

Table 4: Image contextualization (IC) results (%). Loc., Mot., and M stand for Location, Motivation, and METEOR, respectively.

	📍	🗺	🛰	🗺 & 🛰
InternVL-2.5-8B	**62.47**	25.00	53.12	20.53
Qwen2.5-VL-7B	65.19	58.17	**66.59**	64.20
Llama-3.2-11B	40.00	**56.30**	44.50	48.04
GPT4o-mini	**67.61**	49.95	61.50	65.19
Gemini-1.5-Flash	**80.83**	74.43	67.52	65.62

Table 5: Location verification (LV) F1 scores (%). 📍 is the candidate location as a text, while 🗺 and 🛰 are the corresponding map and satellite views.

tions presented in three formats: map view, satellite view, or both. As a baseline, we also evaluate MLLM performance when candidate locations are provided in text form. As shown in Table 5, F1 is often lower with aerial views compared to the text baseline, in particular for commercial models. This suggests that many MLLMs can recognize the location but struggle to match the image to its corresponding aerial view. Model behaviors further diverge by modality: Llama-3.2-11B performs worse than random (49.18%) when given the candidate location as text, while Qwen2.5-VL-7B achieves its best scores with satellite imagery. Gemini-1.5-Flash consistently leads across all settings, and GPT4o-mini is the only model to benefit from combining map and satellite views. Despite these differences, overall F1 scores remain modest, underscoring the difficulty of this multi-image reasoning task. Error cases in Figure 6 illustrate common failure modes: the dam example shows the model's difficulty in distinguishing fine-grained structural details, while the aqueduct example highlights limitations of using a fixed zoom level (15), as discussed further in Appendix H.

Verdict prediction Figures 4d compare models under both *multilingual* and *English-only balanced* settings. Proprietary models show strong linguistic robustness, maintaining comparable performance between multilingual and English claims, regardless of RIS evidence availability. By contrast, open-weight models display much larger disparities across languages. InternVL2.5-8B and Llama-3.2-11B, in particular, perform markedly worse on multilingual claims, with Llama-3.2-11B suffering the steepest drop. Interestingly, Qwen2.5-VL-7B shows the opposite trend, achieving slightly better results in multilingual settings. Across all systems, providing RIS evidence consistently improves results, and the effect is most pronounced for models struggling with multilingual inputs. These results highlight substantial variation in cross-lingual generalization: while leading proprietary models achieve near-linguistic parity, many open-weight alternatives remain challenged in multilingual contexts. Overall performance levels also underscore the difficulty of verdict prediction: no model exceeds 80% F1. The *English-only* performance of GPT4o-mini with DEFAME or with RIS evidence does not differ much. This suggests that RIS constitutes the most useful source of evidence, and other evidence search tools introduced by the DEFAME framework, like web search, image search, and geolocation, have a limited impact on performance.

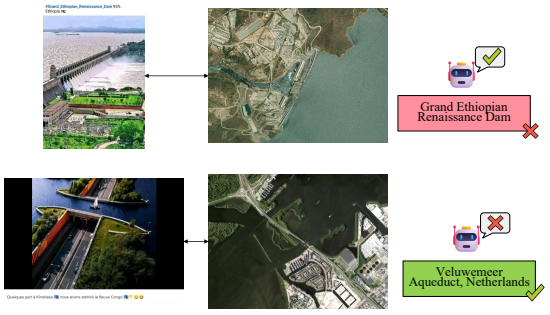


Figure 6: Location verification error examples with Gemini-1.5-Flash on M4FC test. Satellite images are sourced from ESRI World Imagery.

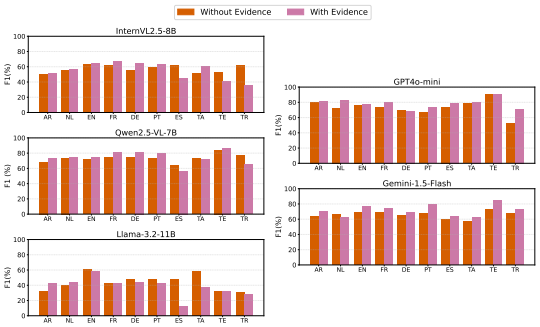


Figure 7: Verdict prediction (VP) F1 scores (%) across different languages. AR: Arabic, NL: Dutch, EN: English, FR: French, DE: German, PT: Portuguese, ES: Spanish, TA: Tamil, TE: Telugu, TR: Turkish.

Intermediate Tasks						GPT4o-mini				Qwen2.5-VL-7B			
VCE	CIP	FD	IC	LV	ER	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
✓						60.92	72.41	71.41	71.91	54.84	66.66	71.05	68.79
✓						53.62	73.14	53.40	61.73	49.26	64.52	58.80	61.53
✓	✓					60.92	72.61	70.98	71.78	53.20	66.03	62.73	64.34
✓		✓				61.31	73.85	70.89	72.34	51.43	65.13	60.76	62.87
✓			✓			62.58	74.14	73.20	73.67	54.96	67.79	69.98	68.87
✓			✓	✓		62.66	74.21	73.89	74.05	55.10	68.63	69.88	69.25
✓					✓	68.82	78.13	77.06	77.59	63.30	72.97	75.61	74.27
✓	✓	✓	✓	✓	✓	68.88	79.50	77.06	78.26	63.42	73.03	76.39	74.67
✓	✓	✓	✓	✓	✓	56.28	74.34	54.51	62.90	53.76	72.63	57.44	64.15
✓	✓					61.48	72.03	73.59	72.80	55.69	69.46	68.74	69.10
✓		✓				58.61	73.61	66.20	69.71	52.58	66.34	63.48	64.88
✓			✓			84.90	**91.82**	86.10	88.87	**78.21**	**85.79**	**82.56**	**84.14**
✓			✓	✓		84.79	91.74	86.03	88.79	78.16	85.78	82.49	84.10
✓	✓	✓	✓	✓	✓	**85.98**	91.38	**87.04**	**89.16**	76.12	**85.79**	81.42	83.55

Table 6: Verdict prediction (VP) results on M4FC test set (*English-only*, *imbalanced*) by combining intermediate tasks and retrieved evidence (%). ER: RIS evidence retrieval. ✓ and ✓ indicate the inclusion of predicted and ground truth intermediate task outputs and evidence, respectively.

To further analyze *multilingual* performance, We disaggregate results by language in Figure 7. Proprietary models again lead overall, generally benefiting from RIS evidence. GPT4o-mini performs weakest on Turkish, while Gemini underperforms on Spanish and Tamil. Telugu is the language on which models perform best, surpassing several higher-resource ones. Yet, open-weight models exhibit smaller or inconsistent gains from evidence, and sometimes perform worse when evidence is added, with Telugu ranking among their weakest languages. These findings suggest that cross-lingual generalization is highly model-dependent, underscoring the need of multilingual evaluation.

Verdict prediction with intermediate tasks Table 6 presents a detailed evaluation of how intermediate task outputs affects verdict prediction. We consider both predicted and ground truth intermediate outputs, in the *imbalanced*, *English-only* setting. We use GPT4o Mini and Qwen2.5-VL-7B, the best performing MLLMs for verdict prediction (see Figures 4d). A case study is shown in Appendix I. Replacing ground truth claims with extracted ones substantially decreases F1, showing that inaccuracies in extracted claims harm downstream verdict prediction. Adding all or some of the other intermediate predicted outputs and the web evidence improves F1. Web evidence has the strongest impact. Surprisingly, using only predicted image contextualization, despite it relying on the same web evidence, performs worse than using the web evidence directly. This suggests that many errors are introduced during image contextualization, and that the web evidence is not sufficient to predict most context items during that intermediate step. When ground truth intermediate outputs are available, image contextualization yields the largest performance boost, improving F1 by 16.96 percentage points. In that case, adding web evidence provides little to no improvement. The highest F1 is achieved by combining the web evidence and all ground truth intermediate tasks, highlighting the complementarity of the intermediate tasks towards verdict prediction. These results show the promises of combining AFC tasks and open the door for future work on how to best implement these combinations, in terms of accuracy, interpretability, and computational efficiency.

5 Conclusion

We introduce M4FC, a real-world dataset for multimodal fact-checking. The images in M4FC have a broad geographic coverage, representing diverse cultures and world events. The dataset supports 10 claim languages and spans six tasks, including two new ones, broadening the scope of multimodal AFC. Our baseline results show that M4FC is a challenging dataset for MLLMs and SOTA AFC methods. Furthermore, we showed that intermediate tasks can have a high impact on downstream verdict prediction performance.

Limitations

We identify four limitations to M4FC.

First, M4FC does not cover videos and audio. Videos have recently become the most prevalent misinformation modality verified by fact-checkers (Dufour et al., 2024), due to the rise of social media platforms like TikTok. However, it is challenging to obtain the videos used in real-world multimodal claims. The fact-checking articles usually only include a screenshot, and the videos themselves are often removed from the social media platforms where they appeared. Implementing a data collection pipeline to create a real-world dataset like M4FC, but with videos instead of images, is an important direction for future work.

Second, despite the creation of scraping scripts customized for each fact-checking organization, a lot of manual verification was needed. For example, many image URLs had to be corrected. We also did not cover fact-checking web pages that use JavaScript. This limits the possibilities to automatically and easily scale the dataset. Future work should consider better automated ways to scrape the raw content of fact-checking web pages.

Third, while the images were fact-checked by human experts, the corresponding labels were extracted automatically using GPT4o. As a result, some labels may contain errors. However, evaluations conducted by Prolific annotators indicate that labeling errors introduced by GPT4o are limited.

Fourth, the *balanced* and *multilingual* verdict prediction settings may suffer from biases. To avoid relying on news articles to source true claims, we generated additional true claims with GPT4o based on the context items of the image. This introduces some biases: the generated true claims are all formulated in a similar way and they are often more detailed than false claims, e.g., they contain specific dates. A trained verdict prediction classifier might exploit these biases. Hence, we do not recommend training on the *balanced* setting for verdict prediction, but rather using only its test set, after training on the *imbalanced* train set or any other dataset. Furthermore, it is not recommended to train a classifier on the *multilingual* and *imbalanced* setting, as most true claims are contained within two languages, English and Portuguese, while other languages only contain false claims, like Arabic. This verdict imbalance by language might also be exploited as a shortcut by a trained classifier.

Ethics statement

Intended use Research in multimodal AFC tackles the important problem of countering misinformation. By providing a new comprehensive resource, this work enables the development of more performant AFC models. M4FC also contributes to the inclusion of low-resource languages and underrepresented cultural contexts in AFC research. M4FC is intended to be used only for the purpose of multimodal AFC research.

Data access We collected data from publicly available fact-checking articles. Many images show real-world events, including graphic scenes of violence. Furthermore, the images show real people, some of them identified by their names in the annotations. Throughout the annotation process, we did not contact fact-checkers and claimants. While fact-checkers have often hidden account names and profile pictures from images, they did not do it for all of them. We did not take any additional anonymization steps. For these reasons, we do not directly provide the images. Instead, we provide URLs directing to the images and a script to download them. We also provide a script to download the satellite images and maps needed for the location verification task. We follow the policy of the AVeriTeC dataset (Schlichtkrull et al., 2023): we will remove the annotations for an image-claim pair upon requests of the claimants, the people shown with the image, or the authors of the fact-checking article. The annotations of M4FC are released under a CC BY-SA-4.0 license. The code is released under an Apache 2.0 license.

AI assistants use AI assistants were used to correct grammar mistakes and typos.

Acknowledgments

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center (Grant Number: LOEWE/1/12/519/03/05.001(0016)/72) and by the German Federal Ministry of Research, Technology and Space and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We thank Max Glockner, Shivam Sharma, and Chen Liu for their feedback on a draft of this work.

References

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

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. COSMOS: catching out-of-context image misuse using self-supervised learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14084–14092. AAAI Press.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. Defame: Dynamic evidence-based fact-checking with multimodal experts. In *Forty-second International Conference on Machine Learning*.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web. *ArXiv preprint*, abs/2505.17978.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2026–2039, Online. Association for Computational Linguistics.

Dafilab. 2025. Ai image detector. `https://huggingface.co/Dafilab/ai-image-detector`. Accessed: 2025-04-13.

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

Gemini-Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *ArXiv preprint*, abs/2403.03627.

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

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 795–816, New York, NY, USA. Association for Computing Machinery.

Sohail Ahmed Khan, Laurence Dierickx, Jan-Gunnar Furuly, Henrik Brattli Vold, Rano Tahseen, Carl-Gustav Linden, and Duc-Tien Dang-Nguyen. 2024. Debunking war information disorder: A case study in assessing the use of multimedia verification tools. *Journal of the Association for Information Science and Technology*.

Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas L. Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, and Duc-Tien Dang-Nguyen. 2023. Visual user-generated content verification in journalism: An overview. *IEEE Access*, 11:6748–6769.

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

Tsung-Yi Lin, Serge Belongie, and James Hays. 2013. Cross-view image geolocalization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *13th International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

MetaAI. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Mikedata. 2025. Real vs fake image model vit base. `https://huggingface.co/mikedata/real_vs_fake_image_model_vit_base`. Accessed: 2025-04-13.

Annique Mossou and Ross Higgins. 2021. A beginner's guide to social media verification. Accessed: 2023-09-15.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 16–25, New York, NY, USA. Association for Computing Machinery.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2025. Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5041–5050.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Kha-Luan Pham, Minh-Khoi Nguyen-Nhat, Anh-Huy Dinh, Quang-Tri Le, Manh-Thien Nguyen, Anh-Duy Tran, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. 2024. Ookpik- a collection of out-of-context image-caption pairs. In *MultiMedia Modeling*, pages 132–144, Cham. Springer Nature Switzerland.

Prithiv. 2024. Deep-fake-detector-model. `https://huggingface.co/prithivMLmods/Deep-Fake-Detector-Model`. Accessed: 2025-04-13.

Qwen-Team. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*. ERIC.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1337–1345.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. 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 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and grounding multimodal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574.

Craig Silverman. 2013. Verification handbook. Accessed: 2023-09-15.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "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*, pages 7845–7864, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Tonglet, Gabriel Thiem, and Iryna Gurevych. 2025. COVE: COntext and VEracity prediction for out-of-context images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2029–2049, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiaying Wu, Fanxiao Li, Min-Yen Kan, and Bryan Hooi. 2025. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. *ArXiv preprint*, abs/2505.15489.

Yuzhuo Xiao, Zeyu Han, Yuhan Wang, and Huaizu Jiang. 2025. Xfacta: Contemporary, real-world dataset and evaluation for multimodal misinformation detection with multimodal llms. *ArXiv preprint*, abs/2508.09999.

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

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

A Labeling prompts

Table 7 presents the prompt instructions provided to GPT4o to extract task labels and metadata from the fact-checking articles.

Table 8 shows the prompt instructions provided to GPT4o for the post-processing steps.

B Data collection and labeling details

Data scraping Following the approach of Tonglet et al. (2024), we retrieved all URLs archived on the Wayback Machine[7] under each organization's web domain. At the time of data collection, all included organizations were current members of the IFCN or under review for membership renewal. We retained URLs containing keywords such as *photo*, *image*, or *picture*, translated into the relevant language. From the raw HTML files, we extracted the article title, publication date, text content, and the URL of the fact-checked image using custom Python scripts. All image URLs were manually verified and corrected when necessary. We manually removed false positives (e.g., non-multimodal claims or fact-checking tutorials), articles containing fact-checking annotations on the image (e.g. an image overlaid with a red cross by the fact-checkers to indicate that it is fake), and duplicate image-claim pairs accross organizations. The final dataset includes 4,982 articles.

GPT4o labeling and post-processing All GPT4o outputs are generated in English and follow a predefined dictionary format. Task labels and metadata that are not available in the article are annotated as "not enough information". Additional post-processing steps with GPT4o include: (1) normalizing all dates to the YYYY-MM-DD format, (2) rewriting claims to remove any language revealing the verdict, and (3) extracting the claim in its original language in the article and identifying that language. Instances are taken into account for a given task if GPT4o successfully extracts the corresponding labels. All verdicts are mapped to the following categories:

[7] archive.org/help/wayback_api.php

Please extract the following information from the debunking article and provide the response in the exact format below. Each item should be in one row, and if any information is unavailable, return "None". Do not output anything else.

Verification strategy: [What verification strategies were used to check the misinformation? (e.g., reverse image search, keyword search, geolocation, deepfake detection, satellite imagery, street view, etc.)]

Verification tools: [What tools were utilized for verification? (e.g., Google, Yandex, Google Earth, etc.)]

Claim: [What is the claim accompanying the image?]

Claimant: [Who made the claim?]

Claimant's Intent: [What is the claimant's intent or agenda behind making the claim?]

Language of the FC article: [In which language is the fact-checking article written?]

Language of the claim: [In which language is the claim accompanying the image written?]

Gold evidence URLs: [List all URLs that link to external webpage evidence.]

Gold image evidence URLs: [List all URLs linking to image evidence.]

Claimed date: [According to the claim, when was the image taken?]

Claimed location: [According to the claim, where was the image taken?]

Claimed people: [Who are the people shown in the image, according to the claim?]

Claimed things: [What objects, buildings, plants, or animals are depicted in the image, according to the claim?]

Claimed event: [What event is being shown in the image, according to the claim?]

Was the image used before? (Provenance): [Did the fact-checkers retrieve a previous version of the same image? (Answer with Yes, No, or Unknown)]

Source: [Which person or organization first took or published the image, according to the fact-check? This is different from the claimant who published the claim with the image.]

Date: [When was the image taken?]

Location: [Where was the image actually taken?]

Motivation: [Why was the image taken?]

People: [Who is shown in the actual image?]

Things: [What objects, buildings, plants, or animals are shown in the actual image?]

Event: [What event is depicted in the actual image?]

Here is the debunking article: `{Article}`

Table 7: GPT4o prompt template to extract task labels and metadata from fact-checking articles.

Post-processing step: Normalize dates to YYYY-MM-DD format
Prompt: Normalize the following dates to YYYY if it is a year, to YYYY-MM if it is a year and month, or YYYY-MM-DD if it is a specific day.
If there is more than one date in the text, repeat the operation for each date, separating them with ";" .
Answer only with the normalized date.
Date: `{DATE}`
Normalized date:

Post-processing step: Normalizing claims
Prompt: Normalize the following claim about an image by keeping only the actual content.
Remove information about the claimant, unnecessary information such as "the image shows", and any words indicating the veracity of the claim, such as "viral", "accurately", "purported".
Keep all information that is part of the claim itself.
Answer only with the normalized claim text.
Claim: `{CLAIM}`
Normalized claim:

Post-processing step: Extracting claim in the original language
Prompt: You are given a fact-checking article debunking a claim, and the translation of this claim in English. Provide the claim as it appears in its original language in the article text. Remove information about the claimant, unnecessary information such as "the image shows", and any words indicating the veracity of the claim, such as "viral", "accurately", "purported" (in the original language).
Answer only with the claim in the original language.
Fact-checking article: `{ARTICLE}`
Claim: `{CLAIM}`
Claim in the original language:

Post-processing step: Identify original language
Prompt: In which language is the following claim written. Answer only in one word with the corresponding language.
Claim: `{CLAIM}`
Language:

Post-processing step: Generate true claims for out-of-context images
Prompt: You are given context information describing an image. Combine this information in one caption of maximum 30 words.
Write the facts only, avoid journalistic style and adjectives, avoid introducing new information.
Date: `{DATE}`
Location: `{LOCATION}`
People: `{PEOPLE}`
Event: `{EVENT}`
Caption:

Table 8: GPT4o prompt templates to post-process M4FC data.

Open the URL and read the Fact-Checking article to verify the labels proposed below.

URL: https://youturn.in/factcheck/100years-before-pandemic-images.html

Claim: People wearing masks during the Spanish Flu pandemic 100 years ago.
Claim in the original language: 100 ஆண்டுகளுக்கு முன்பு ஸ்பானிஷ் ஃப்ளூ பெருந்தொற்றின் போது மாஸ்க் அணிந்து இருந்த மக்கள்.
Verdict: false (out-of-context)
Claimant: Social media users
Claimant motivation: To draw parallels between the COVID-19 pandemic and the Spanish Flu pandemic
Claimed date: 1918-1920
Claimed location: Various locations during the Spanish Flu pandemic
Claimed people: People wearing masks
Claimed things: Masks, protective gear
Claimed event: Spanish Flu pandemic
Source: Country Living, Flickr, AP Images, Alamy
Date: 1939; 1953; 1913
Location: Paris
Motivation: To protect against a snowstorm; To protect against air pollution; Fashion photography
People: Women in protective gear; Muriel Bush and Ruth Newyer; Women in modern attire
Things: Plastic face coverings; War surplus gas masks; Fashion masks
Event: Snowstorm; Air pollution; Fashion photography

Figure 8: An example instance provided to the Prolific annotators.

true, false, mixture, unverified. The latter two, which account for only 27 instances, are excluded from verdict prediction. False claims are further categorized into: false (out-of-context), false (manipulated/fake),[8] and false (AI-generated).

Generating synthetic true claims For each authentic image labeled as false (out-of-context), we generate a synthetic true claim using GPT4o, combining the following context items, if at least one is available: date, location, people, and event. In the *balanced* verdict setting, we replace the original false claim with the corresponding synthetic true claim in half of the instances where a true claim was successfully generated. The prompt is provided in Appendix A.

Selecting images for location verification Images suitable for location verification are identified semi-automatically. First, we use keyword-based filtering to find articles referencing maps, satellite imagery, or detailed geographic verification (e.g., at the street or landmark level). We then manually verify that these images show sufficient background detail to be compared with maps or satellite views, yielding a final set of 65 images. Because the dataset includes only accurate location labels by default, we collect two inaccurate candidates per instance. GPT4o generates a list of likely locations for the image, and the first two with valid GeoNames[9] coordinates are selected as inaccurate candidate locations.

Temporal dataset split The dataset is split temporally into train (60%), dev (10%), and test (30%) sets, using cutoff dates of 2022-09-22, 2023-01-18, and 2024-09-01, respectively, to avoid temporal leakage (Glockner et al., 2022).

C Data validation - Prolific study

We recruited annotators on the platform Prolific to validate the quality of the data extracted by GPT4o from fact-checking articles. Annotators were paid at the default rate of £9 per hour. We applied the following criteria to select two annotators per study: they need to (1) be fluent speaker of English and their primary language needs to be the one used by the fact-checking organization, (2) have no language-related disorders, (3) have an approval rate of 99% and (4) have already participated in 50 annotation tasks in the past.

C.1 Annotation guidelines

We used the Potato annotation tool (Pei et al., 2022) to design the data validation interface. Each instance consists of the URL of the fact-checking arti-

[8] Excluding AI-generated images.

[9] geonames.org

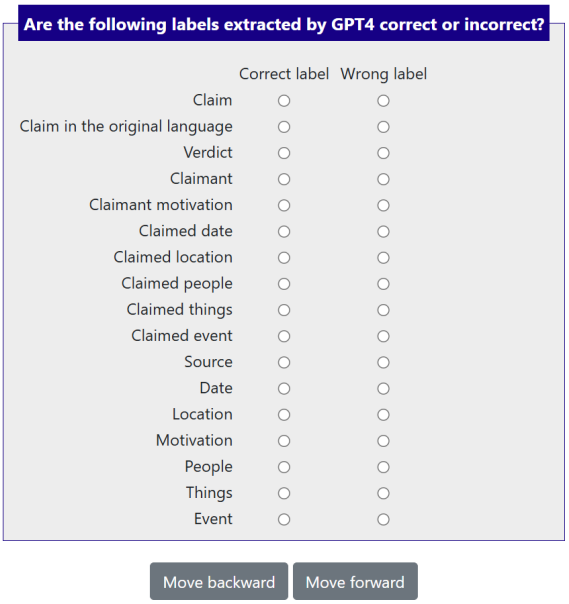


Figure 9: The annotators' interface to validate the labels extracted by GPT4o from fact-checking articles.

cle and the labels extracted by GPT4o, as shown in Figure 8. For each label, annotators have to select either "correct label" or "wrong label", as shown in Figure 9. Cases where the label is "not enough information" and the annotator selects "wrong label" are automatically mapped to "missing label", while others are mapped to "incorrect label".

C.2 Percentage agreement and IAA scores

Table 9 provides the detailed percentages of correct, incorrect, and missing labels, and the IAA scores computed with Randolph's κ. Results are provided per type of label and averaged over all 23 Prolific studies. These results exclude the attention tests. For the percentage of correct, incorrect, and missing labels, we report both the lower bound, where a label is considered correct only if both annotators agree on it, and the upper bound, where an answer is considered correct if at least one annotator agrees on it. IAA is reported both with and without taking into account the cases where the label is "not enough information".

D Additional dataset statistics

Task combination distribution Figure 10 presents the distribution of task combinations in the dataset. Each row corresponds to a task, and each column represents a unique combination of tasks. The bar chart above indicates the number of instances associated with each combination. The majority of instances are annotated with labels for at least four tasks, while combinations involving only one or two tasks are rare.

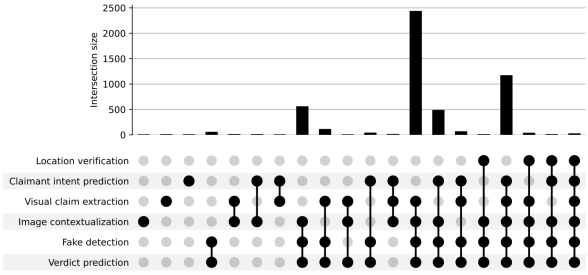


Figure 10: Number of M4FC instances with labels for different task combinations.

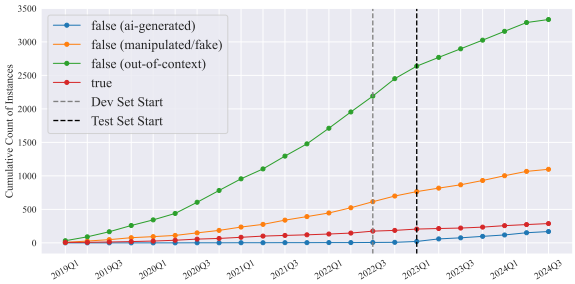


Figure 11: Cumulative distribution of the verdicts over time. The categories with few instances, *unverified*, *mixture*, *false (authentic)*, are not displayed.

Verdict distribution In the *imbalanced* setting of verdict prediction, true claims represent only 5.9% of the entire data. This proportion increases to 26.0% in the *balanced* setting. False claims are dominated by out-of-context misinformation (72%), aligning with prior findings (Dufour et al., 2024). Manipulated and man-made fake images account for 24.0%, while only 3.6% of the false claims involve AI-generated images. Furthermore, AI-generated images mostly appeared from 2023, and are therefore concentrated in the test set. Figure 11 shows the cumulative evolution over time of the verdict labels, with the cutoff dates for the dev and test set indicated by vertical dashed lines.

Claim language distribution The language distribution under the *multilingual* setting is as follows: English (63.8%), Arabic (12.4%), Portuguese (5.0%), French (4.6%), Spanish (4.0%), German (3.6%), Telugu (3.0%), Dutch (1.6%), Tamil (1.0%), Turkish (1.0%).

Geographic distribution Figure 12 compares the geographic distribution of images with location context items in M4FC and two previous datasets with known location labels, VERITE and 5Pils.

Label	Correct (%)		Incorrect (%)		Missing (%)		Randolph's κ	
	Low	Up	Low	Up	Low	Up	w/o NEI	w/ NEI
Claim (English)	83.2	98.8	16.8	1.2	0.0	0.0	0.70	0.70
Claim (original language)	66.7	90.6	21.9	1.1	11.5	8.3	0.52	0.53
Verdict	89.6	99.3	10.4	0.7	0.0	0.0	0.81	0.81
Claimant	82.0	98.4	12.2	0.5	5.8	1.2	0.74	0.67
Claimant motivation	73.3	98.4	6.5	0.9	20.3	1.2	0.78	0.51
Claimed date	49.6	86.2	7.4	1.5	43.0	12.3	0.62	0.32
Claimed location	79.6	97.6	9.0	0.5	11.4	1.9	0.75	0.65
Claimed people	85.0	98.3	4.8	0.0	10.1	1.7	0.83	0.72
Claimed things	77.2	94.7	5.3	0.9	17.5	4.4	0.85	0.65
Claimed event	81.2	98.1	8.9	1.0	9.9	1.0	0.80	0.67
Source	63.7	93.9	20.0	2.9	16.4	3.2	0.45	0.37
Date	64.0	97.4	20.9	0.8	15.1	1.8	0.51	0.41
Location	76.4	97.3	11.3	0.2	12.3	2.4	0.67	0.57
Motivation	74.8	99.3	12.6	0.0	12.6	0.7	0.66	0.56
People	79.1	99.3	8.5	0.0	12.4	0.7	0.75	0.61
Things	80.6	97.1	12.1	0.7	13.6	2.2	0.82	0.64
Event	74.3	96.6	11.0	0.9	13.3	2.5	0.67	0.56
Total	75.8	96.6	11.0	0.9	13.3	2.5	-	-

Table 9: Detailed results of the Prolific data validation study. Task labels and metadata are highlighted in blue and orange, respectively. Low and Up indicates the lower and upper bound scores, respectively.

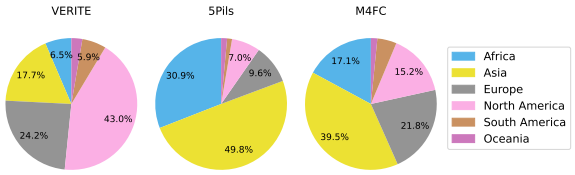


Figure 12: Distribution by continent of the images with known location of VERITE, 5Pils, and M4FC.

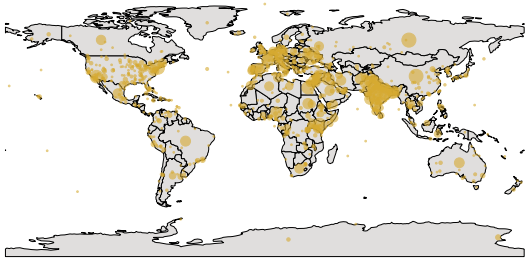


Figure 13: Geographic distribution of images in M4FC with known location.

VERITE is skewed toward Western contexts, while 5Pils is largely concentrated in Asia and Africa. In contrast, M4FC shows the highest geographic diversity, featuring images from a broad range of cultural contexts. Figure 13 further highlights the geographic diversity of M4FC. However, some areas remain underrepresented, such as East Asia, South America, and Oceania.

Geographic distribution per organization Figures 14 and 15 show the geographic distribution by country of the claimed and real location of M4FC images, respectively. One map is drawn for each fact-checking organization in M4FC. The maps show that fact-checking organizations tend to focus on images claimed to be associated with their own country. For example, Factly.in and Dfrac.org, both based in India, focus on images with claimed location India. Notable exceptions include Fatabyyano.net, Knack.be, and Dogrula.org. In addition to their own country, many organizations verify images with Ukraine as a claimed location, which reflects the ongoing conflict that started in 2022. For some organizations, the real locations at the country level do not differ significantly from the claimed locations. This is the case for organizations like Factly.in and Indiatoday.in, where all claimed and real locations are within India. This

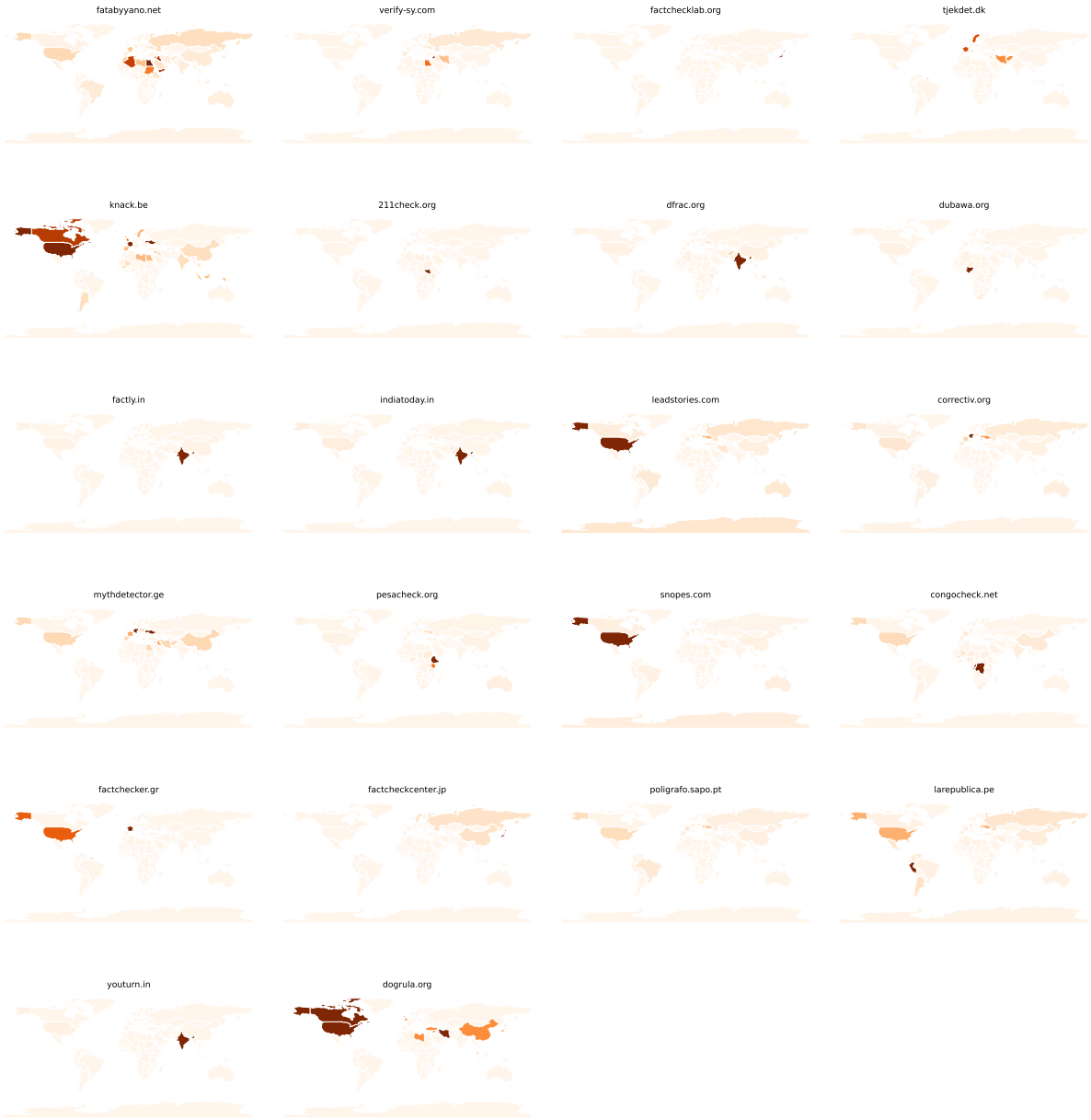


Figure 14: Distribution by country and by fact-checking organization of the claimed locations of images in M4FC.

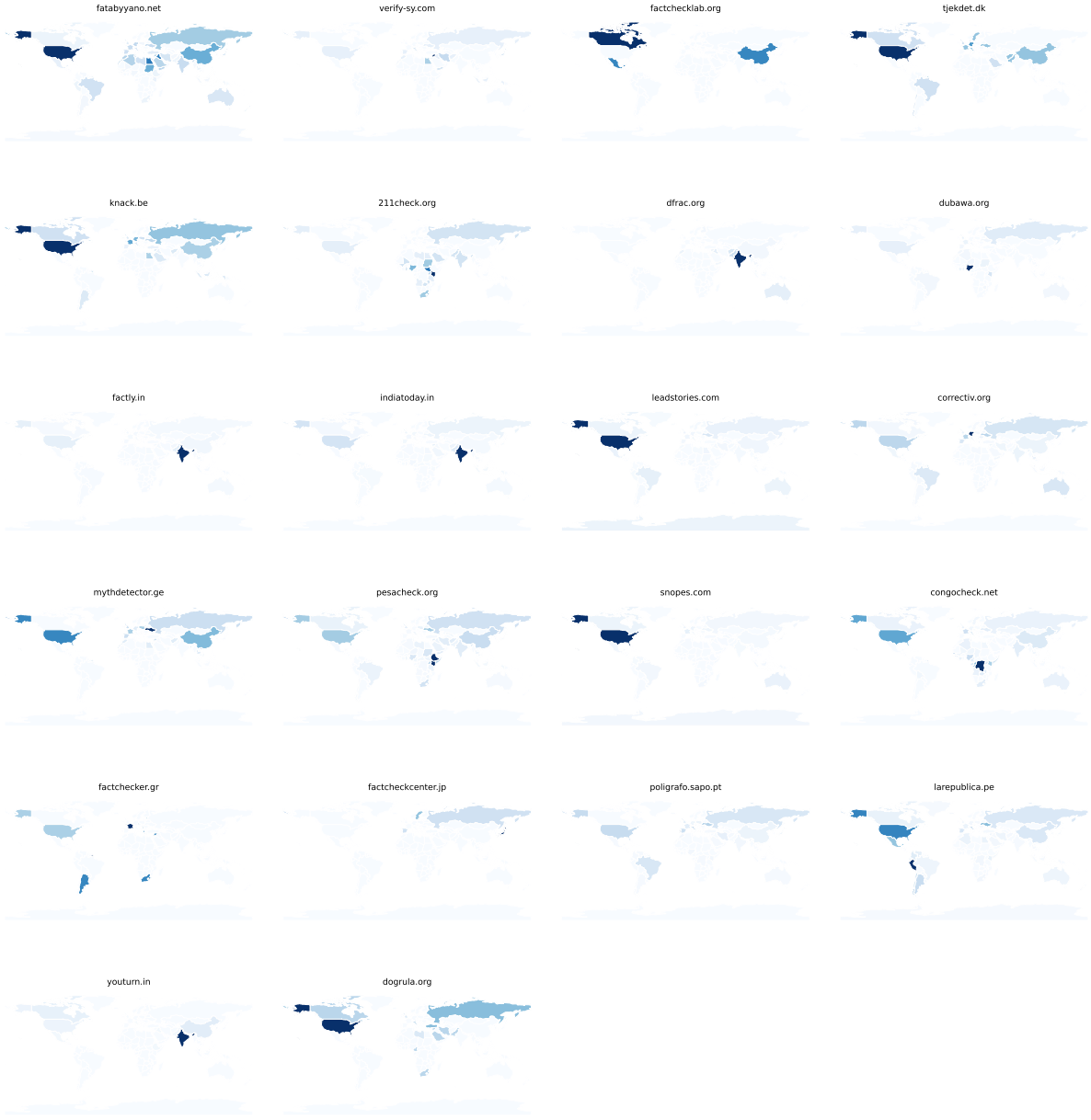


Figure 15: Distribution by country and by fact-checking organization of the locations of images in M4FC.

Task: Visual Claim Extraction
Prompt: You are an assistant designed to support fact-checking. Your task is to help identify the primary claim conveyed by the shared image that requires verification. Clearly articulate this claim in a single, concise, and neutral sentence or short paragraph, beginning with "Claim:"

Task: Claimant Intent Prediction
Prompt: You are a helpful assistant supporting fact-checking. Given the image and the related claim: `{claim}`, your task is to help initialize the prediction of the user's motivation/intention for sharing this image for further analysis. Clearly articulate it in a single, concise, and neutral sentence or short paragraph, starting with "Motivation: To"

Task: Image Contextualization
Prompt: You are an AI assistant designed to help analyze information. Given the image and related article `{evidence}`, your task is to identify the following important information:

1. People: Who is shown in the image?
2. Things: Which animal, plant, building, or object are shown in the image?
3. Event: Which event is depicted in the image?
4. Date: When was the image taken?
5. Location: Where was the image taken?
6. Motivation: Why was the image taken?
7. Source: Who is the source of the image?

Please provide your answers in the following format in English:
People: [Answer],
Things: [Answer],
Event: [Answer],
Date: [Answer],
Location: [Answer],
Motivation: [Answer],
Source: [Answer]
If there is no answer, just leave it blank, like an empty string. Do not output anything else. The output must be in English.

Task: Location verification
Prompt: You are a helpful assistant designed to support fact-checking. Your task is to verify whether a candidate location is accurate for an image. You are given the image, and [a map | a satellite | a map and a satellite] view of the candidate location as input. Answer only with true or false.

Task: Verdict prediction Prediction (optionally with text evidence)
Prompt: You are an AI assistant designed to evaluate the veracity of multimodal claims. Given the claim, the related image, and retrieved evidence, your task is to analyze the provided information and predict the more likely veracity: True or False. Please do not refuse or respond with "not enough information." Your prediction will assist professional fact-checkers in further analysis.
Please provide your prediction as "Veracity: True" or "Veracity: False".
In one or more paragraphs, output your reasoning steps. In the separate final line, output your prediction. (Please don't output anything else except the Veracity Prediction)
Claim: `{CLAIM}`
Retrieved Evidences: `{EVIDENCE}`
Let's think step by step.

Table 10: Prompt templates for the baseline experiments.

	Rouge-L	BERTScore	METEOR
InternVL2.5-8B	25.70	51.43	18.04
Qwen2.5-VL-7B	28.44	58.21	20.09
Llama-3.2-11B	19.82	48.73	12.01
GPT4o-mini	26.94	56.2	18.70
Gemini-1.5-Flash	**32.78**	**64.7**	**23.67**

Table 11: Visual claim extraction results (%).

	Rouge-L	BERTScore	METEOR
InternVL2.5-8B	15.30	32.20	8.04
Qwen2.5-VL-7B	15.35	34.81	7.93
Llama-3.2-11B	11.76	32.04	5.43
GPT4o-mini	11.14	30.23	5.47
Gemini-1.5-Flash	**17.13**	**45.72**	**8.70**

Table 12: Claimant intent prediction results (%).

Model	Acc	F1	Pre	Rec
Deep-Fake Detector ❄	58.05	70.31	64.05	77.92
Deep-Fake Detector 🔥	**69.49**	78.88	70.58	89.38
AI-Image Detector ❄	36.74	6.41	56.14	3.40
AI-Image Detector 🔥	68.34	78.49	69.21	**90.66**
RealFake Detector ❄	55.48	60.17	70.00	52.76
RealFake Detector 🔥	67.87	**79.64**	**76.16**	83.47

Table 13: Fake detection results (%). ❄ and 🔥 indicate predictions of the base pre-trained model and of the model fine-tuned on M4FC train set, respectively.

could either mean that location is less frequently misrepresented in out-of-context claims in India, or that the difference in location is at a more fine-grained level, e.g., different cities.

Web evidence statistics After filtering, we obtain web pages with the Google RIS engine for only 48.62% of the images in M4FC. In total, 9,693 web pages are retrieved, with an average of 4 per image, excluding images with 0 web pages. The web pages are written in 84 different languages. English is the most frequent (65%), followed by Spanish (5%), French (3%), and Arabic (2%).

E Prompt templates for the M4FC tasks

Table 10 reports the prompt templates used for the baseline experiments.

F Implementation details

MLLMs All models are evaluated in one zero-shot run on one A100 GPU with temperature set to 0. We use the HuggingFace (Wolf et al., 2020) implementation of the open-weight models. GPT4o-mini and Gemini-1.5-Flash are accessed via the OpenAI API and Google Cloud API, respectively.

Fake detectors fine-tuning We fine-tune all models for fake detection with the Adam optimizer and a learning rate of $2e-5$ for 10 epochs.

DEFAME We employ the Google Custom Search API for both text and image retrieval. We allow up to three iterations, with each retrieval step returning the top-3 evidence items. The maximum output length of a single LLM generation is limited to 2,000 tokens (`max_tokens`), while the overall fact-checking process, including the final report and all intermediate steps, is capped at 64,000 tokens (`max_result_len`).

COVE We reproduce all steps of the COVE pipeline, except the collection of evidence via direct search, which has limited impact on real-world misinformation detection (Tonglet et al., 2025).

RIS evidence For each image, we retrieve up to 20 web pages with the Google RIS engine. We extract their textual content using Trafilatura (Barbaresi, 2021).[10] To prevent evidence leakage, we exclude any web pages published by a curated list of fact-checking organizations, as well as those published after the date of the fact-checking article from which the claim was sourced. For tasks that require text evidence, we treat each web page retrieved with RIS as a single evidence. All 20 text evidence are provided in the prompt, sorted by cosine similarity with the image using CLIP embeddings (Radford et al., 2021).

G Additional Results

Tables 11,12, 13, and 14 show the results tables on M4FC test set for Visual claim extraction, claimant intent prediction, fake detection, and verdict prediction in the *multilingual balanced* setting. These tables correspond to the results in Figures 4 and 7.

H Location verification at different zoom levels

As shown in the second error example of Figure 6, better results could be obtained at different zoom levels, as illustrated in Figure 16. However, the appropriate zoom level may vary a lot. Future work should consider directly incorporating the desired zoom level for the aerial views as a parameter of their modeling approach. We provide preliminary

[10] trafilatura.readthedocs.io

Language	GPT4o-mini		Gemini-1.5-Flash		Qwen2.5-VL-7B		InternVL2.5-8B		LLaMA-3.2-11B	
	w/o evi	w/ evi	w/o evi	w/ evi	w/o evi	w/ evi	w/o evi	w/ evi	w/o evi	w/ evi
Arabic	80.11	81.09	63.11	70.13	68.36	73.06	50.52	52.41	31.67	42.57
Dutch	72.56	82.43	65.73	62.95	73.48	74.68	55.62	57.37	40.21	44.54
English	76.12	78.14	68.44	76.34	71.52	74.58	64.12	65.19	60.68	58.32
French	73.25	80.04	68.28	74.28	73.89	81.25	61.78	67.67	43.26	43.29
German	69.07	68.27	64.34	69.52	74.47	81.36	55.93	64.40	47.83	44.36
Portuguese	66.73	73.42	67.12	79.10	72.71	79.30	59.65	63.84	47.23	43.24
Spanish	73.86	78.87	60.26	64.17	63.40	56.55	62.56	44.85	47.69	12.31
Tamil	78.37	80.22	57.55	62.32	72.92	71.65	52.20	60.32	57.82	37.40
Telugu	90.32	90.91	72.48	84.23	84.05	86.73	52.60	41.21	31.59	32.78
Turkish	52.16	71.49	67.28	72.62	76.76	65.37	62.10	35.98	30.20	28.08

Table 14: Verdict prediction (*multilingual*, *balanced*) F1 per language, with (w/) and without (w/o) evidence (%).

Zoom			F1
10	✓		68.85
10		✓	66.59
10	✓	✓	64.14
15	✓		**74.43**
15		✓	67.52
15	✓	✓	65.62
17	✓		73.63
17		✓	70.55
17	✓	✓	69.75

Table 15: Location verification F1 at different zoom level on M4FC test set with Gemini-1.5-Flash (%).

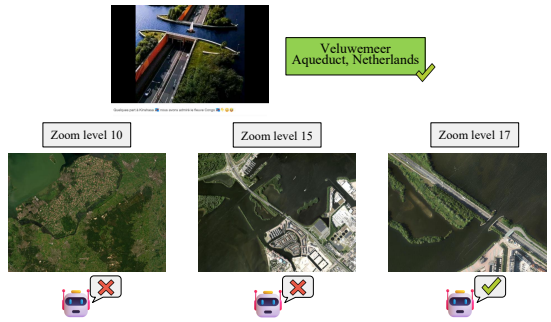


Figure 16: Illustration of location verification with satellite views at different zoom level. Satellite images are sourced from ESRI World Imagery.

results in Table 15, showing the F1 of Gemini-1.5-Flash on M4FC test set at three different zoom levels: 10, 15, and 17. The overall best performance is obtained with a zoom level of 15.

I Pipeline case studies

Figure 17 shows an example of an automated pipeline experiment where the tasks are performed sequentially, using the output of intermediate tasks as input for subsequent ones. The false (out-of-context) claim is the same as in Figure 1. The pipeline starts with claim extraction, which provides a claim that reflects the original social media post well. The intent of the claimant is also accurately interpreted. Based on the social media post, it appears that the claimant is expressing support for Donald Trump. In parallel, the image is recognized as authentic. The output of the next task, image contextualization, is of lower quality. Many errors are introduced, such as the people involved, the event shown, and the date of the picture. The location is correct, although limited to the country level, preventing the step of location verification. However, despite the errors introduced during image contextualization, it is still possible to provide the correct verdict.

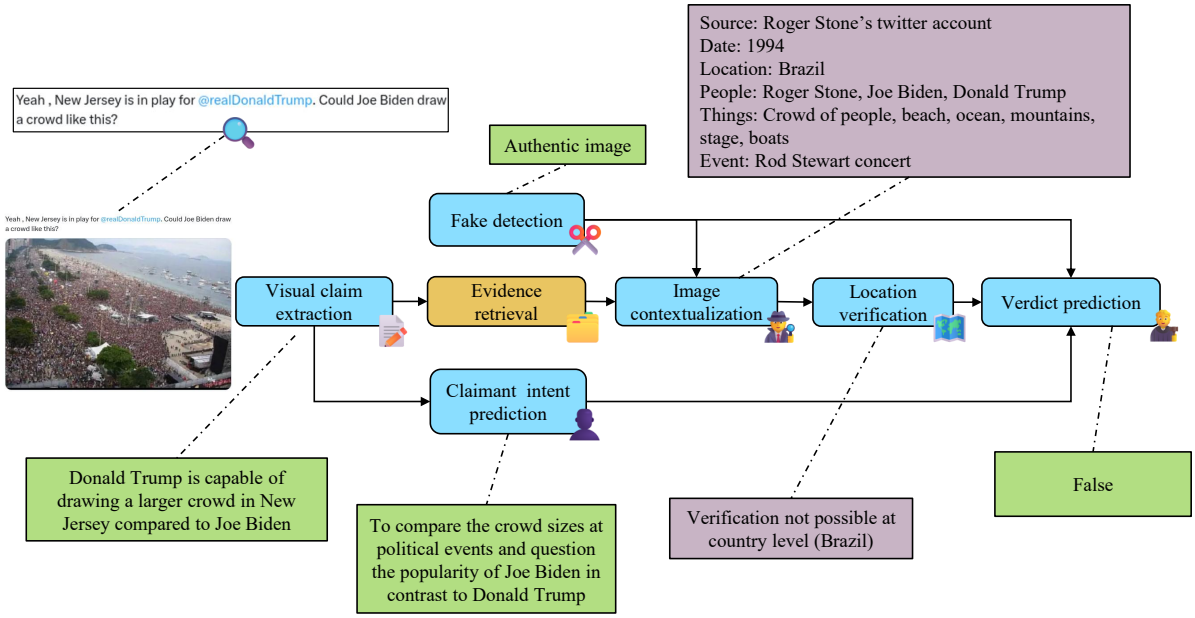


Figure 17: An example of an experiment with intermediate outputs. The instance is the same as in Figure 1. Purple boxes indicate (partially) incorrect answers, while green boxes indicate correct answers.