VERITAS: The First Dynamic Benchmark for Multimodal Automated Fact-Checking

Anonymous ACL submission

Figure 1: Example claims from the VERITAS benchmark, including media, claim date, and claim integrity score. The two claims on the right showcase lower-level annotations of media/claim properties used to infer the overall integrity. Each annotation contains a justification. Claims are added quarterly via an automated pipeline.

Abstract

The growing scale of online misinformation urgently demands Automated Fact-Checking (AFC). Existing benchmarks for evaluating AFC systems, however, are largely limited in terms of task scope, modalities, domain, language diversity, realism, or coverage of misinformation types. Critically, they are *static*, thus subject to data leakage as their claims enter the pretraining corpora of LLMs. As a result, benchmark performance no longer reliably reflects the actual ability to verify claims. We introduce *Verified Theses and Statements* (VERITAS), the first *dynamic* benchmark for multimodal AFC, designed to remain robust under ongoing large-scale pretraining of foundation models. VERITAS currently comprises 24,000 real-world claims from 108 professional fact-checking organizations across 54 languages, covering textual and audiovisual content. Claims are added *quarterly* via a fully automated seven-stage pipeline that normalizes claim formulation, retrieves original media, and maps heterogeneous expert verdicts to a novel, standardized, and disentangled scoring scheme with textual justifications. Through human evaluation, we demonstrate that the automated annotations closely match human judgments. We commit to updating VERITAS in the future, establishing a leakage-resistant benchmark, supporting meaningful AFC evaluation in the era of rapidly evolving foundation models. We will make the code and data publicly available.

1 Introduction

Mis- and disinformation lead the current short-term risks for global society (Elsner et al., 2025). Humans perceive multimodal information as more credible than text alone (Newman et al., 2012), making images and videos powerful tools for persuasion (Hameleers et al., 2020; Greifeneder et al., 2020). As a result, multimodal content achieves higher engagement and spreads faster, particularly in the context of misinformation (Li and Xie, 2020; Zannettou et al., 2018; Wang et al., 2021). In their analysis, Dufour et al. (2024) found that visual misinformation accounts for up to 80% of fact-checked claims. At the same time, generative methods like NANO BANANA (Google, 2025b) or SORA 2 (OpenAI, 2025b) evolve rapidly, producing realistic deepfakes, outpacing countermeasures. Fact-checking, however, is a very laborious task (Hassan et al., 2015), which gives rise to robust *multimodal Automated Fact-Checking (AFC)*.

To evaluate AFC methods, numerous benchmarks have been proposed. However, many lack full modality coverage, missing images or videos (Schlichtkrull et al., 2023; Cao et al., 2025; Geng et al., 2025), restrict the task scope to isolated subproblems (Papadopoulos et al., 2024; Tonglet et al., 2024), or rely heavily on synthetically generated claims (Xu et al., 2024a, 2025), raising concerns about realism. Additionally, practically all AFC benchmarks model the task as a single-label classification problem, conflating orthogonal claim properties into coarse labels, often ignoring uncertainty and finer-grained properties, and treating severe confusions (*True* vs. *False*) equally to less severe ones (e.g., *True* vs. *Not Enough Information*).

More fundamentally, all existing AFC benchmarks are *static* and therefore vulnerable to *data leakage*. State-of-the-art AFC systems rely on large language models (LLMs) that are continually pretrained on public web data. Since benchmark claims originate from the same public sources, models may implicitly encode both claims and verdicts in their parametric knowledge, undermining meaningful evaluation. We empirically substantiate this effect in Figure 4, demonstrating a sharp performance drop on claims published after models' knowledge cutoff date.

To address these limitations, we introduce *Verified Theses and Statements (*VERITAS*)*, the first *dynamic* AFC benchmark. VERITAS is extended quarterly with newly emerging real-world claims, effectively mitigating data leakage in future splits through fully automated data acquisition. It is also the first benchmark to combine real-world multimodal claims featuring both images and videos. Unlike the vast majority of previous work, VERITAS also provides multilingual, open-domain, and balanced claim data with verdict annotations accompanied by textual justifications for the rulings, derived from expert ground truth. Figure 1 shows example claims from VERITAS.

Moreover, we contribute a novel rating scheme that (1) disentangles previously coarse-grained rulings into separate fine-grained claim and media properties, (2) incorporates uncertainty by modeling each of these properties on a scale from -1 to 1 allowing to (3) employ Mean Squared Error (MSE) as a evaluation metric to reward near-correct predictions while strongly penalizing flipped decisions.

Finally, we benchmark state-of-the-art multimodal AFC systems and strong LLM baselines, revealing substantial room for improvement. We will publish the benchmark data and pipeline code online, and we commit to extending VERITAS every coming quarter until at least Q4 2028.

2 Related Work

Benchmarks by Task Scope and Domain Automated Fact-Checking (AFC) remains a challenging and unsolved problem, particularly in realistic, open-domain settings (Akhtar et al., 2023; Dmonte et al., 2025; Schlichtkrull et al., 2024). To tackle the complexity, most benchmarks restrict the task scope, either by decomposing AFC into sub-tasks, such as image contextualization (Tonglet et al., 2024), out-of-context detection (Xu et al., 2024b; Papadopoulos et al., 2024; Luo et al., 2021; Tonglet et al., 2025; Aneja et al., 2021), manipulation detection (Shao et al., 2023), check-worthiness estimation (van der Meer et al., 2025), fact-check retrieval (Papadopoulos et al., 2025), content interpretation (Jin et al., 2024), deepfake detection (Skoularikis et al., 2025), claim disambiguation (Staliunaite and Vlachos, 2025; Glockner et al., 2024), claim detection (Cheema et al., 2022), claim normalization (Sundriyal et al., 2023), or claim matching (Pisarevskaya and Zubiaga, 2025). Some benchmarks limit the domain to specific platforms like Reddit (Nakamura et al., 2020), or topics like elections (Khatiwada et al., 2025), finance (Rangapur et al., 2025), climate charts (Su et al., 2025), or the Ukraine-Russia war (Bondielli et al., 2024). In contrast, VERITAS targets the holistic and open-domain, end-to-end task of **claim verification**, where the goal is to predict expert ratings.

Benchmarks by Modality and Annotation Early AFC benchmarks are predominantly text-only, including FEVER (Thorne et al., 2018) and LIAR (Wang, 2017), with later publications such

Benchmark		Primary task	Images	Videos	Dynamic	Automated	Real claims	Misinfo coverage	# Languages	Open-domain	Justifications	Balanced	# Instances	Annotation
M³A	(Xu et al., 2024a)	CV	✓	✓	-	✓	-	-	1	News	-	-	7.3 M	4 Classes
MDAM³-DB	(Xu et al., 2025)	CV	✓	✓	-	✓	-	-	1	News	-	-	90 K	6 Classes
XFACTA	(Xiao et al., 2025)	CV	✓	-	◑	-	✓	◑	1	SM	-	✓	2.4 K	4 Classes
MUMIN	(Nielsen and McConville, 2022)	CV	✓	-	-	✓	✓	✓	41	✓	-	-	13 K	3 Classes
KHATIWADA ET AL.	(Khatiwada et al., 2025)	CV	✓	-	-	✓	✓	◑	1	SM	-	-	77 K	5 Labels
CLAIMREVIEW2024+	(Braun et al., 2025)	CV	✓	-	-	-	✓	✓	1	✓	-	-	300	4 Classes
M4FC	(Geng et al., 2025)	CV	✓	-	-	-	✓	✓	10	✓	-	✓	7.0 K	4 Classes
REALFACTBENCH	(Yang et al., 2025b)	CV	✓	-	-	-	✓	✓	1	✓	-	-	6.0 K	2 Classes
MR^2	(Hu et al., 2023)	CV	✓	-	-	-	✓	◑	2	✓	-	-	15 K	3 Classes
VLDBENCH	(Raza et al., 2025)	CV	✓	-	-	-	✓	◑	1	News	-	✓	63 K	2 Classes
AVERIMATEC	(Cao et al., 2025)	CV	✓	-	-	-	✓	-	1	✓	✓	-	1.3 K	4 Classes
FACTIFY	(Mishra et al., 2022)	CV	✓	-	-	-	◑	◑	1	✓	-	✓	50 K	5 Classes
FACTIFY 3M	(Chakraborty et al., 2023)	CV	✓	-	-	-	◑	◑	1	✓	-	-	3 M	5 Classes
MMFAKEBENCH	(Liu et al., 2025)	CV	✓	-	-	-	-	◑	1	✓	-	-	11 K	3 Classes
OMNIFAKE	(Li et al., 2025a)	CV	✓	-	-	-	-	-	1	SM	-	-	127 K	3 Classes
TRUE	(Niu et al., 2025)	CV	-	✓	-	✓	✓	✓	1	✓	✓	-	2.9 K	2 Classes
GROUNDLIE360	(Yang et al., 2025a)	CV	-	✓	-	-	✓	✓	1	✓	-	✓	2.0 K	6 Classes
MMOOC	(Xu et al., 2024b)	OOCD	✓	✓	-	✓	-	-	1	News	-	-	455 K	2 Classes
VERITE	(Papadopoulos et al., 2024)	OOCD	✓	-	-	✓	◑	-	1	✓	-	✓	1.0 K	3 Classes
NEWSCLIPPINGS	(Luo et al., 2021)	OOCD	✓	-	-	✓	◑	-	1	News	-	✓	988 K	2 Classes
5PILS-OOC	(Tonglet et al., 2025)	OOCD	✓	-	-	-	✓	-	1	✓	-	✓	1.2 K	2 Classes
COSMOS (test split)	(Aneja et al., 2021)	OOCD	✓	-	-	-	◑	-	1	News	-	✓	1.7 K	2 Classes
DGM^4	(Shao et al., 2023)	MD	✓	-	-	✓	-	✓	1	News	-	-	230 K	2 Classes
VERITAS (Ours)		CV	✓	✓	✓	✓	✓	✓	54+	✓	✓	✓	24 K+	5 Scores

Table 1: Overview of related multimodal AFC benchmarks. **Primary task**: CV = Claim Verification, OOCD = Out-of-Context Detection, MD = Manipulation Detection. **Dynamic**: Whether the authors highlight the benchmark as extensible with new claims (◑) and if the authors committed to update it regularly (✓). **Automated**: Whether the construction pipeline is fully autonomous. **Real claims**: Whether the benchmark is entirely made from real-world claims (and perhaps close derivations) (✓) or contains some (◑) or mostly (-) synthetic, i.e., invented claims. **Misinfo coverage**: Whether the claims cover the full (✓) or broad (◑) spectrum of contemporary misinformation types for the given modalities, or just selected types (-). **Open-domain**: SM = Social Media. **Justifications**: Whether the benchmark contains explanations in addition to the annotations.

as AVERITEC (Schlichtkrull et al., 2023) addressing previous issues like temporal leakage (Glockner et al., 2022). However, only multimodal benchmarks can capture the important role of visual content—see Tab. 1 for a detailed comparison. We consider multimodal AFC benchmarks to incorporate claims with associated images and/or videos, unlike datasets that use multimodal evidence but retain text-only claims as proposed by Tang et al. (2024); Yao et al. (2023); Wang et al. (2025). Multimodal AFC benchmarks are often limited in size (Zlatkova et al., 2019), or miss images or videos entirely. Only a few datasets support both images and videos (Xu et al., 2024a, 2025). In contrast, VERITAS covers *all* common modalities. Additionally, from all multimodal AFC benchmarks, only a few provide justifications for explanation (Cao et al., 2025; Niu et al., 2025). Almost all benchmarks use a single-label n-class annotation scheme, often entangling different, orthogonal properties (such as media authenticity and claim veracity) in the same coarse category with unclear separation. The only exception is Khatiwada et al. (2025) who employ a multi-labeling scheme. VERITAS explicitly models claim and media properties with disentangled, uncertainty-aware scores and textual justifications.

Dynamic Claim Datasets Dynamic benchmarking has mainly been explored in adversarial settings where data collection and model training co-evolve (Shirali et al., 2023), most prominently in DYNABENCH (Kiela et al., 2021). By contrast, VERITAS is updated independently of evaluated models. Sustaining a dynamic dataset requires *automation*: Many benchmarks rely on human annotation (Xiao et al., 2025) or synthetic claim generation (Xu et al., 2024a, 2025), raising concerns about scalability or realism. MUMIN (Nielsen and McConville, 2022) and TRUE (Niu et al., 2025) are the only ones to accomplish both realism and automation. However, MUMIN lacks videos, and TRUE lacks images. Two dynamic claim repositories exist: CLAIMSKG (Tchechmedjiev et al., 2019) and CIMPLEKG (Burel et al., 2025). Both aggregate fact-checked claims daily, but are text-only and not designed as evaluation benchmarks. Xiao et al. (2025) express intent to "continuously update" their benchmark, XFACTA, yet the most recent change is more than

5 months ago. Fatahi Bayat et al. (2025) propose a dynamic yet text-only benchmark, consisting not of claims but of prompts of different complexities to probe the factuality of LLMs.

In a nutshell, and as can be seen in Tab. 1, prior work always lacks multiple important properties. In contrast, VERITAS combines real-world multimodal claims, fine-grained annotations derived from expert-provided rulings with justifications, and a fully automated construction pipeline. Its dynamic nature makes VERITAS the first benchmark specifically designed for long-term, leakage-resistant evaluation of multimodal AFC systems.

3 The VERITAS Construction Pipeline

VERITAS (*Verified Theses and Statements*) is a multimodal and multilingual benchmark for evaluating Multimodal Automated Fact-Checking (MAFC) systems. It consists of real-world claims combining text, images, and videos, and is designed as a *dynamic* benchmark that is extended quarterly. VERITAS is constructed via a seven-stage pipeline enabling large-scale, standardized claim acquisition and annotation.

LLMs have proven to be reliable enough for automated annotation (Khatiwada et al., 2025). Across stages, we employ LLMs from the GPT 5 (OpenAI, 2025a) and GEMINI 2.5 (Google, 2025a) families as these represent the state of the art in multimodal language modeling at the time of data collection, achieving leading performance in image and video understanding tasks, respectively. We select model sizes based on task complexity (see App. A for LLM version numbers). For multi-step reasoning, we apply chain-of-thought prompting; for generation tasks, we use few-shot in-context examples to constrain output style. Please refer to the supplementary material for the prompts. Figure 2 summarizes the seven stages we are going to describe next—App. B contains additional details.

3.1 Stage 1: Review Discovery

Goal. Retrieve expert-annotated ground truth claims. We collect fact-check reviews via the ClaimReview[1] schema and, where missing, we determine the language of each claim. **Output.** Parsed ClaimReview data of about 371 K reviews published between Jan 1, 2016 and Dec 31, 2025, including claim text, rating, review URL, date, language, and appearance URLs.

[1]claimreviewproject.com

Figure 2: The seven stages of VERITAS repeated on a quarterly basis.

3.2 Stage 2: Publisher Identification

Goal. Ensure review credibility. We identified 832 distinct publishers via review URLs, and retain reviews only if they originate from credible fact-checkers (cf. App. B), dismissing about 55 K (15%) of the reviews not meeting this criterion. **Output.** Reviews from professional fact-checking organizations meeting international standards.

3.3 Stage 3: Article Scraping

Goal. Obtain full fact-check context. Since ClaimReview provides no justifications for the ratings, we scrape the original fact-checking articles incl. media. With GPT-5 NANO we extract the main textual body via span detection, preserving the original text but removing cookie notices, ads, and other UI noise. We discard 3.3 K (0.9%) paywalled and 6.2 K (1.7%) otherwise inaccessible articles as well as 7.0 K (1.9%) unrealistically short articles. **Output.** Reviews with cleaned article content.

3.4 Stage 4: Appearance Retrieval

Goal. Recover original claim sources and media. An *appearance* denotes the original location where a claim surfaced (e.g., a social media post). ClaimReview provides appearance URLs in only 13.3% of cases. For reviews with missing appearances, we prompt GPT-5 MINI to extract appearance URLs from the article text, including archived versions (e.g., Perma.cc; see App. D.2), which are resolved to original sources when possible. If original content cannot be downloaded, archived versions are used. We retain up to two successfully scraped appearances, omitting videos longer than 5 minutes or larger than $128\,MB$ to maintain a reasonable claim scope. $98\,K$ (26.6%) Reviews without any valid appearance are discarded. **Output.** $80\,K$ Reviews with one or two downloaded appearances.

3.5 Stage 5: Claim Normalization

The raw claim text provided by ClaimReview is often malformed, e.g., exposing the verdict, constituting incomplete sentences, or including unnecessary meta information like "Social media posts claiming..." and omitting media completely. **Goal.** Produce precise, self-contained claims with relevant media.

Media Filtering. We retain only media inherent to the claim, removing duplicates via cosine similarity larger than 0.85 in the embedding space of `Qdrant/clip-ViT-B-32-vision`, an image-specialized CLIP variant. Videos are inspected with GEMINI 2.5 FLASH, images with GPT-5 MINI, to exclude illustrative or editorial content. Media precedence follows: original appearance > archived appearance > article media. At most four media items are kept per claim, prioritizing videos.

Claim Reformulation. Using GEMINI 2.5 PRO and GPT-5.2 for video claims and non-video claims, respectively, raw claim texts are rewritten into about $59\,K$ concise, self-contained statements, conditioned on article content, appearances, and metadata (e.g., claimant and date). Reformulated claims explicitly reference associated media when present and preserve the original language.

Claim Validation. Roughly $17\,K$ (28.0%) claims are discarded by LLM validation with GPT-5 MINI and GEMINI 2.5 FLASH as they (i) expose the verdict, (ii) lack required media, or (iii) exceed 70 words to limit claim scope. **Output.** $43\,K$ validated, well-formed claims with associated media if relevant.

3.6 Stage 6: Verdict Standardization

Goal. Standardize heterogeneous fact-check ratings. Fact-checking organizations employ heterogeneous rating schemes that often conflate multiple

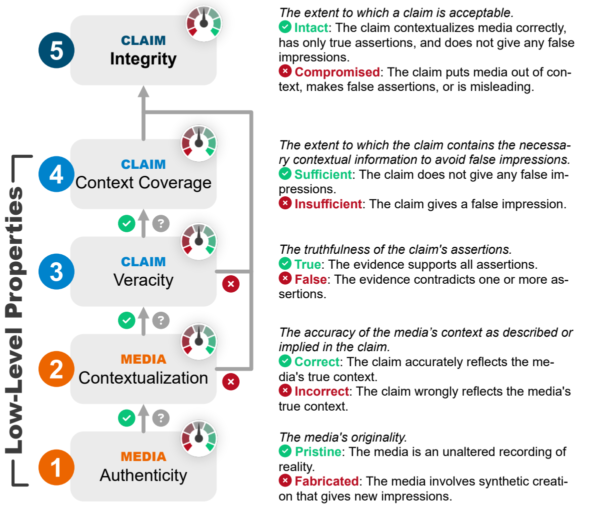


Figure 3: Verdict derivation (Stage 6), assessing properties (1) to (4). Negative decisions in (2) or (3) result in early termination. Definitions are shown on the right.

dimensions, using labels such as “half true,” “legend,” or “three Pinocchios” (see Fig. 7). To obtain a consistent and interpretable representation, we decompose claim assessment into four independent properties that are evaluated sequentially (Fig. 3): (1) *Media Authenticity*, (2) *Media Contextualization*, (3) *Veracity*, and (4) *Context Coverage*. To retain a single scalar notion of overall acceptability, we further introduce a higher-order property termed *Integrity* (5). Integrity captures whether a claim is acceptable as presented: Low integrity indicates that at least one lower-level property is violated, whereas high integrity corresponds to positive assessments of Properties (2–4). Integrity constitutes the primary evaluation target, while the lower-level properties serve as explanatory factors that localize the source of compromise and simplify the annotation task. Media-specific properties (1–2) are assessed per media item; if no media is present, evaluation begins at veracity, targeting the textual part. The *Context Coverage* property explicitly captures claims that are technically true yet misleading due to omitted crucial context, a phenomenon previously referred to as cherry-picking (Schlichtkrull et al., 2023).

Inspired by Glockner et al. (2024), who propose to avoid a rigid class-like categorization of claims to incorporate uncertainty, we model each property on a scale from -1 to $+1$, where 0 denotes full uncertainty (? **NEI**, Not Enough Information). A score of $< -1/3$ is considered a “negative” decision, analogously for “positive.” The integrity score is determined directly by the worst score obtained for properties 2 to 4, which we refer to as the *compromising* property. Consequently, a claim is ✅ Intact if all media (if any) are correctly contextualized, the claim veracity is true, and no false impressions arise from missing context; otherwise, it is ❌ Compromised. If a prediction at step (2) or (3) results in a negative decision, the follow-up properties will not significantly impact the integrity. Thus, we terminate evaluation for the claim early for negative decisions at steps (2) and (3).

To increase robustness, we use an ensemble of four LLMs (GPT-5.2, GEMINI 2.5 PRO, CLAUDE SONNET 4.5 (Anthropic, 2025), and LLAMA 4 MAVERICK (Meta AI, 2025), aggregating predictions by their mean. We discard $1.7\,K$ (2.8%) claims that received an ensemble prediction with an internal score difference of more than 1 to keep only instances with high inter-model agreement. Additionally, each ensemble member provides a one-paragraph justification for their decision, reciting the core arguments from the fact-checking article. We instruct GPT-5 MINI to summarize the four justifications into a single one. The justification for the rating of the compromising property serves as justification for the integrity. **Output.** $41\,K$ Claims with high-agreement verdicts and justifications.

3.7 Stage 7: Claim Rectification

Goal. Balance intact and compromised claims. Fact-checkers primarily verify compromised claims since these are the most harmful, cf. Fig. 14a. Only $\sim 0.4\%$ of ClaimReviews yield intact claims. Thus, we generate about $27\,K$ "corrected" (intact) versions of the compromised claims using GPT-5.2 and GEMINI 2.5 PRO, guided by the justification of the compromising property, retaining media.

Consistency between corrected text and media is validated using GPT-5 MINI or GEMINI 2.5 FLASH. To avoid stylistic shortcuts, we use the same LLMs to validate if rectified claims are *shareable*, i.e., relevant, not overly specific, and worth sharing, like original claims. This way, we identify $5.7\,K$ (21.2%) rectified claims for exclusion. Finally, rectified claims are reevaluated using Stage 6. Claims with integrity score $> 1/3$ are kept and replace their originals; all 643 (2.4%) others are discarded along with $4.7\,K$ (17.3%) claims with insufficient ensemble agreement. **Output.** $13\,K$ consistent, shareable, intact rectified claims.

24,000	100%	Claims
11,953	49.8%	❌ Compromised
94	0.4%	❓ NEI
11,953	49.8%	✅ Intact
11,236	46.8%	W/ media
6,865	28.6%	W/ images
4,620	19.2%	W/ videos
12,764	53.2%	W/o media
18.1	±7.9	Words / Claim

Table 2: VERITAS claim statistics, see also Tab. 6.

4 Results

4.1 VERITAS Statistics

VERITAS currently contains $24\,K$ claims spanning 24 quarterly splits from Q1 2020 to Q4 2025. Each quarter comprises $1\,K$ claims balanced to have an equal number of ✅ Intact and ❌ Compromised instances. Earlier periods are excluded due to insufficient review coverage. The most recent quarter is the only one suitable for reliable benchmarking. In addition to the quarterly splits, we release a budget-friendly *longitudinal split* containing 2,400 balanced claims (100 per quarter), enabling temporal analysis across the full time span. Summary statistics are reported in Tab. 2 and App. D.4.

VERITAS contains 8,377 images and 4,785 videos while it covers 54 languages, with 34 occurring at least 50 times. English is most frequent (40.5%), followed by Spanish (10.2%) and Hindi (5.7%), see Fig. 11c for more languages. Claim appearances are predominantly sourced from Facebook (38.6%) and X/Twitter (20.2%), while all other platforms individually contribute less than 6%. AFP Fact Checking accounts for the largest share of reviewed claims (24.1%), with all other organizations contributing below 6% each.

4.2 Validation through Human Evaluation

We validate the automated annotations from stage 6 (Sec. 3.6) via human evaluation. Each claim is independently annotated by ≥ 2 native or C1-level speakers of the claim's language. Annotators sequentially assess all properties, following the exact same annotation procedure as in stage 6, including scoring and providing written justifications.

We report Mean Squared Error (MSE), Mean Absolute Error (MAE), and accuracy after discretizing scores into three bins. MSE is our primary metric, as it penalizes severe confusions more strongly. MSE ≤ 0.04 indicates very high agreement (App. E). Results are summarized in

Integrity	N	MSE (↓)	MAE (↓)	Acc. (↑)
VERITAS	63	**0.034**	**0.092**	**96.8**
- no filtering	66	0.042	0.105	93.9
- no ensemble	66	0.086	0.170	93.9

Table 3: Human evaluation results over N claims. The scores reflect agreement between human annotation and automated VERITAS annotations for integrity for three different settings. **MSE** = Mean Squared Error, **MAE** = Mean Absolute Error, **Accuracy** = share of matches of 3-bin discretized scores (in %).

Method	Search	MSE (↓)	MAE (↓)	Acc. (↑)
GEMINI 2.0 FLASH	-	0.74	0.71	32.1
GEMINI 2.5 FLASH	-	0.85	0.57	65.9
GEMINI 3 PRO	-	**0.55**	**0.37**	**81.9**
GPT-4O	-	0.65	0.65	36.9
GPT-5.2	-	0.70	0.69	33.5
LLAMA 4 MAVERICK	-	0.97	0.74	41.8
GEMINI 2.0 FLASH	✓	0.73	0.57	58.0
GEMINI 2.5 FLASH	✓	0.68	0.48	71.2
GEMINI 3 PRO	✓	**0.39**	**0.35**	**74.6**
GPT-4O	✓	0.65	0.50	64.2
GPT-5.2	✓	0.45	0.40	70.6
LLAMA 4 MAVERICK	✓	1.04	0.72	49.6
DEFAME (w/ GPT-5.2)	✓	0.55	0.49	60.4
LOKI (w/ GPT-5.2)	✓	0.86	0.59	61.8

Table 4: Model evaluation on the **Q4 2025 split** with(out) search tool (single runs).

Tab. 3, where we also include two ablated VERITAS variants: *no filtering* (keeping claims with high ensemble disagreement) and *no ensemble* (replacing the ensemble with just GPT-5.2). The evaluation yields an MSE below 0.04, corresponding to roughly one flipped prediction among 100 otherwise correct judgments, which we consider very strong agreement. Accuracy appears lower (96.8%) because it penalizes mild disagreements involving ? NEI equally with severe confusions between ✓ Intact and ✕ Compromised. The observations are confirmed by confusion matrices, see Fig. 19. Overall, these results confirm that the automated verdict mapping closely aligns with human judgments, validating the VERITAS construction pipeline. Additionally, the increased error rates for the ablated variants in Tab. 3 indicate the benefits of the ensemble approach and the agreement filtering. See Sec. G for more details.

4.3 Benchmarking AFC Methods on VERITAS latest quarter

We analyze baselines and current AFC methods on the Q4-2025 split to assess performance on the most recent data—particularly after the knowledge cutoff from all tested methods. We consider six foundation models: GPT-5.2 and GEMINI 3 PRO as state-of-the-art multimodal models; GPT-4O, GEMINI 2.5 FLASH, and GEMINI 2.0 FLASH as earlier-generation models to study knowledge-cutoff effects; and LLAMA 4 MAVERICK as a representative open-source model. Each model is evaluated once using parametric knowledge only and once with web search augmentation, see App. F.1 for the tool implementation details.

GEMINI 3 PRO and GEMINI 2.5 FLASH process videos natively. For all other models, we represent videos via five evenly spaced frames. Additionally, we compare against two recent fact-checking systems, DEFAME (Braun et al., 2025) and LOKI (Li et al., 2025b), both with GPT-5.2 as backbone. Tab. 4 displays the results.

Across all baselines, GEMINI 3 PRO achieves the strongest results on all metrics in both evaluation settings, even surpassing the AFC-specialized models DEFAME and LOKI. The second-best model, GPT-5.2, trails by $+0.15$ and $+0.06$ MSE, depending on retrieval access. Open-source LLAMA 4 MAVERICK performs the worst, also under search augmentation, lagging behind even older proprietary models such as GEMINI 2.0 FLASH. This highlights a substantial performance gap between current open-source and proprietary multimodal LLMs for AFC. Surprisingly, despite its specialization, LOKI outperforms only LLAMA.

4.4 The Role of Knowledge Cutoff

We evaluate the same models on the longitudinal VERITAS split. Fig. 4 shows the results. All evaluated models exhibit a marked increase in MSE after their respective Knowledge Cutoff Dates (KCDs). On average, MSE rises from roughly 0.6 to above 0.8, substantially reducing the margin to trivial *always* ? *NEI* behavior. The effect is particularly pronounced for GEMINI 3 PRO, whose MSE increases from below 0.4 to over 0.6. We can rule out the increased share of videos as a confounding factor, since the same phenomenon is observed for text-only claims, see Fig. 17 for the breakdown.

4.5 Real-World Insights

The data obtained from stages 1–6 roughly reflect the natural distribution of real-world fact-checked claims and reveal significant temporal trends, despite the dismissal of numerous reviews and claims. The share of AI-generated media steadily increases

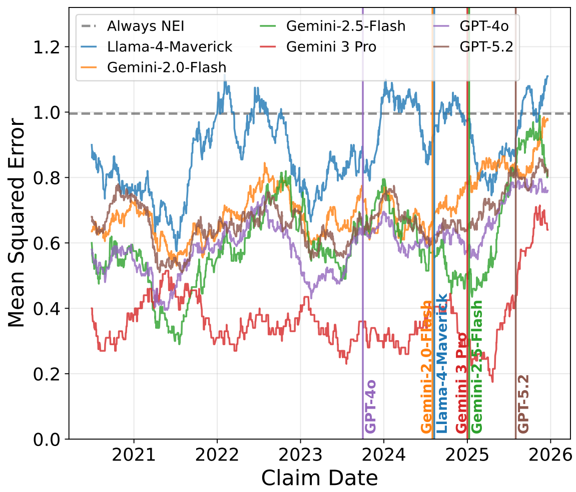


Figure 4: Baseline performance without web search on the **longitudinal split** using a 200-claim moving average window, single runs. Lower is better. Vertical lines indicate knowledge cutoff dates.

over time, exceeding 25% of claims by Q4-2025, while the proportion of pristine media correspondingly declines (Fig. 15). The fraction of claims containing media—particularly videos—also increases over time (Fig. 14). Platform coverage shifts are observable, with a declining share of X/Twitter and a growing share of Instagram (Fig. 15). Finally, the volume of all published ClaimReviews rises until Q1-2025 and decreases thereafter (Fig. 6).

5 Discussion

Data leakage severely undermines static AFC benchmarks. The increase in error rate of models after their KCD provides strong empirical evidence that LLMs partially encode real-world claims and verdicts from before their KCD, even without external retrieval. Since realistic fact-checking concerns newly emerging claims, benchmarks dominated by pre-KCD data systematically overestimate model performance. Notably, the most recent KCD among evaluated models (GPT-5.2: Aug 31, 2025) postdates a large fraction of existing AFC benchmarks, rendering their evaluation increasingly unreliable.

Real-world claims are essential for ethical and realistic evaluation. We argue that high-quality AFC benchmarks must be grounded in real claims. Synthetic claims fail to capture the diversity and realism of real misinformation. Moreover, a generator capable of producing realistic, multimodal misinformation at scale would constitute a severe ethical risk. Consequently, relying on synthetic multimodal misinformation is neither methodologically nor ethically defensible for benchmarking purposes. In contrast to this—and to most other benchmarks generating synthetic claims—stage 7 of VERITAS creates *intact* claims conditioned on *expert ground truth*.

Substantial room for improvement. No evaluated model approaches an MSE of 0.1, which we consider a reasonable threshold for acceptable AFC. Even the best-performing model (GEMINI 2.5 PRO, MSE $= 0.39$) corresponds to approximately one flipped prediction in ten otherwise correct cases. In contrast, LLAMA 4 MAVERICK attains an MSE near 1.0, effectively matching constant **NEI** prediction. These results underscore that multimodal AFC remains far from solved, even with state-of-the-art models and retrieval.

Outlook. The dynamic design of VERITAS will capture emerging misinformation trends through quarterly updates. Future work must address potential paradigm shifts, such as increased prevalence of audio-only content or platform restrictions that limit content access. Scaling claim retrieval beyond the current 24 K claims toward the full set of over 380 K reviews would further improve coverage and analytical power. Finally, ongoing evaluation of newly released foundation models and AFC systems on future splits is needed for a complete and up-to-date performance comparison.

6 Conclusion

We introduced VERITAS, the first dynamic benchmark for multimodal automated fact-checking, designed to remain robust under continual large-scale pretraining of foundation models. By collecting real-world, multilingual, multimodal claims via a fully automated pipeline and updating the benchmark quarterly, VERITAS mitigates data leakage that increasingly invalidates static AFC benchmarks. Human evaluation validates both the annotation scheme and the construction process. Baseline experiments reveal that current state-of-the-art multimodal LLMs fall far short of reliable fact-checking performance, while performance on pre–knowledge-cutoff data substantially overestimates real-world capability. VERITAS establishes a principled foundation for meaningful, future-proof evaluation of multimodal AFC systems.

7 Limitations

Despite its scope and level of automation, VERITAS has several limitations. First, we do not perform rigorous cross-lingual claim deduplication. As a result, semantically equivalent claims may appear multiple times across languages, particularly when fact-checking organizations publish parallel versions of the same claim (e.g., AFP). While this reflects real-world dissemination, it may introduce redundancy within splits.

Second, although rectified claims are generated using targeted prompting and automated validation, their linguistic style is not always optimal. Some rectified claims may remain verbose or insufficiently natural, which could expose exploitable shortcuts for inference.

Third, VERITAS is designed around current modalities, platforms, and publication practices. Potential paradigmatic shifts—such as increased prevalence of audio-only content, emergence of new communication platforms, restrictions on media download, or the discontinuation of fact-checking infrastructures (e.g., ClaimReview by Google or Data Commons)—could impair future data collection. While the pipeline has operated reliably for data from the past six years, its long-term robustness cannot be guaranteed.

Finally, although fully automated, the construction of VERITAS incurs non-trivial API and computational costs. Processing the full benchmark required approximately \$14.9 K in total, corresponding to about \$600 per quarter split. Additionally, hosting LLAMA 4 MAVERICK in-house required roughly 2.7 K GPU hours on 8 NVIDIA H100 GPUs, which may limit reproducibility for resource-constrained labs.

8 Ethical Considerations

With the introduction of VERITAS, we aim to support society in addressing mis- and disinformation by enabling more reliable and meaningful evaluation of multimodal AFC systems. By solving issues of previous benchmarks, VERITAS is intended to accelerate responsible AFC research sustainably.

VERITAS relies exclusively on real-world claims that have already been fact-checked by professional organizations. This design choice mitigates ethical risks associated with *generating* synthetic misinformation, which could otherwise contribute to the creation or dissemination of harmful content. Using expert-derived judgments promotes transparency, accountability, and reproducibility, while preserving the realism necessary for valid evaluation.

At the same time, potential misuse risks exist. Models evaluated on VERITAS could be repurposed for surveillance, censorship, or selective moderation, particularly in politically sensitive contexts. Moreover, although VERITAS reduces evaluation bias caused by data leakage, it may still reflect geographic, linguistic, or institutional biases inherent in the global fact-checking ecosystem.

All data are sourced from publicly available materials; therefore, we did not anonymize individual claims or media. However, since some content may have been removed from the web after our collection, and to avoid automated scraping, access to VERITAS is restricted to eligible researchers via a gated request process on Hugging Face. Finally, the dataset may contain offensive or disturbing content, which we intentionally retain, as such material is an intrinsic part of real-world fact-checking and necessary for faithful evaluation.

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal Automated Fact-Checking: A Survey. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings*. arXiv.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning. *Preprint*, arXiv:2101.06278.

Anthropic. 2025. Claude Sonnet 4.5 System Card. System Card. Accessed on Jan 5, 2026.

Alessandro Bondielli, Pietro Dell'Oglio, Alessandro Lenci, Francesco Marcelloni, and Lucia Passaro. 2024. Dataset for multimodal fake news detection and verification tasks. *Data in Brief*, 54:110440.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 5383–5417. PMLR.

Grégoire Burel, Martino Mensio, Youri Peskine, Raphael Troncy, Paolo Papotti, and Harith Alani. 2025. CimpleKG: A Continuously Updated Knowledge Graph on Misinformation, Factors and Fact-Checks. In *The Semantic Web – ISWC 2024*, pages 97–114, Cham. Springer Nature Switzerland.

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. In *NeurIPS 2025*.

Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W Question-Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15282–15322, Singapore. Association for Computational Linguistics.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-Claims: A Dataset for Multimodal Claim Detection in Social Media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. Claim Verification in the Age of Large Language Models: A Survey. *Preprint*, arXiv:2408.14317.

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. *Preprint*, arXiv:2405.11697.

Mark Elsner, Grace Atkinson, and Saadia Zahidi. 2025. Global Risks Report 2025. Technical report, World Economic Forum.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33090–33110, Vienna, Austria. Association for Computational Linguistics.

Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2025. M4FC: A Multimodal, Multilingual, Multicultural, Multitask Real-World Fact-Checking Dataset. *Preprint*, arXiv:2510.23508.

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.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Transactions of the Association for Computational Linguistics*, pages 1–18.

Google. 2025a. Gemini 3 Pro Model Card. Model Card. Accessed on Jan 5, 2026.

Google. 2025b. Introducing Nano Banana Pro. https://blog.google/technology/ai/nano-banana-pro/. Accessed on Jan 5, 2026.

Rainer Greifeneder, Mariela Jaffe, Eryn Newman, and Norbert Schwarz. 2020. *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*, 1 edition. Routledge, London.

M. Hameleers, T. E. Powell, T. G. L. A. Van Der Meer, and L. Bos. 2020. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37.

Naeemul Hassan, Bill Adair, J. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking.

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, pages 2901–2912, New York, NY, USA. Association for Computing Machinery.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. MM-SOC: Benchmarking Multimodal Large Language Models in Social Media Platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6192–6210, Bangkok, Thailand. Association for Computational Linguistics.

Prerana Khatiwada, Qile Wang, Kenneth E. Barner, and Matthew Louis Mauriello. 2025. Towards a Multimodal Multi-Label Election-Context Repository for Classifying Misinformation. *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*, 2025:26.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Haiyang Li, Yaxiong Wang, Shengeng Tang, Lianwei Wu, Lechao Cheng, and Zhun Zhong. 2025a. Towards Unified Multimodal Misinformation Detection in Social Media: A Benchmark Dataset and Baseline. *Preprint*, arXiv:2509.25991.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025b. Loki: An Open-Source Tool for Fact Verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Yiyi Li and Ying Xie. 2020. 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.

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. *Preprint*, arXiv:2406.08772.

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.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed on Jan 5, 2026.

Shreyash Mishra, S. Suryavardan, Amrit Bhaskar, P. Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, A. Sheth, and Asif Ekbal. 2022. FACTIFY: A Multi-Modal Fact Verification Dataset. In *DE-FACTIFY@AAAI*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Eryn J. Newman, Maryanne Garry, Daniel M. Bernstein, Justin Kantner, and D. Stephen Lindsay. 2012. Non-probative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19(5):969–974.

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, pages 3141–3153, New York, NY, USA. Association for Computing Machinery.

Kaipeng Niu, Danni Xu, Bingjian Yang, Wenxuan Liu, and Zheng Wang. 2025. Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-role Multimodal Model Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28276–28283.

OpenAI. 2025a. GPT-5 System Card. System Card, OpenAI. Accessed on Jan 5, 2026.

OpenAI. 2025b. Sora 2 is here. https://openai.com/index/sora-2/. Accessed on Jan 5, 2026.

Stefanos-Iordanis Papadopoulos, Ivana Beňová, Sebastian Kula, Michal Gregor, George Karantaidis, Tomáš Javůrek, Marián Šimko, and Symeon Papadopoulos. 2025. Multimodal and Multilingual Fact-Checked Article Retrieval. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, pages 1063–1071, New York, NY, USA. Association for Computing Machinery.

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.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and Few-shot Learning with Instruction-following LLMs for Claim Matching in Automated Fact-checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9721–9736, Abu Dhabi, UAE. Association for Computational Linguistics.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. Fin-Fact: A Benchmark Dataset for Multimodal Financial Fact-Checking and Explanation Generation. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 785–788, New York, NY, USA. Association for Computing Machinery.

Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, S. Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and Mubarak Shah. 2025. VLDBench: Vision Language Models Disinformation Detection Benchmark. *Information Fusion Journal*.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The Automated Verification of Textual Claims (AVeriTeC) Shared Task. *FEVER Workshop at EMNLP 2024*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and Grounding Multi-Modal Media Manipulation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Ali Shirali, Rediet Abebe, and Moritz Hardt. 2023. A Theory of Dynamic Benchmarks. In *ICLR 2023*.

Anastasios Skoularikis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2025. 'Humor, Art, or Misinformation?': A Multimodal Dataset for Intent-Aware Synthetic Image Detection. In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, DHOW '25, pages 95–104, New York, NY, USA. Association for Computing Machinery.

Ieva Staliunaite and Andreas Vlachos. 2025. Dis2Dis: Explaining Ambiguity in Fact-Checking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 246–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Ruiran Su, Jiasheng Si, Zhijiang Guo, and Janet B. Pierrehumbert. 2025. ClimateViz: A Benchmark for Statistical Reasoning and Fact Verification on Scientific Charts. *Preprint*, arXiv:2506.08700.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From Chaos to Clarity: Claim Normalization to Empower Fact-Checking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6594–6609, Singapore. Association for Computational Linguistics.

Chia-Wei Tang, Ting-Chih Chen, Kiet A. Nguyen, Kazi Sajeed Mehrab, Alvi Md Ishmam, and Chris Thomas. 2024. M3D: MultiModal MultiDocument Fine-Grained Inconsistency Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22270–22293, Miami, Florida, USA. Association for Computational Linguistics.

Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II*, pages 309–324, Berlin, Heidelberg. Springer-Verlag.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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. *Preprint*, arXiv:2502.01194.

Michiel van der Meer, Pavel Korshunov, Sébastien Marcel, and Lonneke van der Plas. 2025. HintsOfTruth: A Multimodal Checkworthiness Detection Dataset with Real and Synthetic Claims. *Preprint*, arXiv:2502.11753.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2025. Piecing It All Together: Verifying Multi-Hop Multimodal Claims. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7453–7469, Abu Dhabi, UAE. Association for Computational Linguistics.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuping Wang, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Use of Fauxtography on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:776–786.

Yuzhuo Xiao, Zeyu Han, Yuhan Wang, and Huaizu Jiang. 2025. XFacta: Contemporary, Real-World Dataset and Evaluation for Multimodal Misinformation Detection with Multimodal LLMs. *Preprint*, arXiv:2508.09999.

Qingzheng Xu, Huiqiang Chen, Heming Du, Hu Zhang, Szymon Łukasik, Tianqing Zhu, and Xin Yu. 2024a. M3A: A multimodal misinformation dataset for media authenticity analysis. *Computer Vision and Image Understanding*, 249:104205.

Qingzheng Xu, Heming Du, Huiqiang Chen, Bo Liu, and Xin Yu. 2024b. MMOOC: A Multimodal Misinformation Dataset for Out-of-Context News Analysis. In *Information Security and Privacy*, pages 444–459, Singapore. Springer Nature.

Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025. MDAM3: A Misinformation Detection and Analysis Framework for Multitype Multimodal Media. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 5285–5296, New York, NY, USA. Association for Computing Machinery.

Bingjian Yang, Danni Xu, Kaipeng Niu, Wenxuan Liu, Zheng Wang, and Mohan Kankanhalli. 2025a. A New Dataset and Benchmark for Grounding Multimodal Misinformation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 12571–12577, New York, NY, USA. Association for Computing Machinery.

Shuo Yang, Yuqin Dai, Guoqing Wang, Xinran Zheng, Jinfeng Xu, Jinze Li, Zhenzhe Ying, Weiqiang Wang, and Edith C. H. Ngai. 2025b. RealFactBench: A Benchmark for Evaluating Large Language Models in Real-World Fact-Checking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 13435–13441, New York, NY, USA. Association for Computing Machinery.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. 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*, pages 2733–2743.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 188–202, New York, NY, USA. Association for Computing Machinery.

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 LLM Glossary

Table 5 summarizes the LLMs used in the VERITAS construction pipeline. At the time of pipeline processing, rate limits were too restrictive for the newly introduced GEMINI 3 PRO, therefore we defaulted to GEMINI 2.5 PRO.

LLM name	Stages	Version specifier
GPT-5.2	5, 6	`gpt-5.2-2025-12-11`
GPT-5 MINI	4 – 7	`gpt-5-mini-2025-08-07`
GPT-5 NANO	3	`gpt-5-nano-2025-08-07`
GEMINI 2.5 PRO	5, 6, 7	`gemini-2.5-pro`
GEMINI 2.5 FLASH	5, 7	`gemini-2.5-flash`
CLAUDE SONNET 4.5	6	`claude-sonnet-4-5-20250929`
LLAMA 4 MAVERICK	6	`Llama-4-Maverick-17B-128E-Instruct-FP8`

Table 5: The Large Language Models (LLMs) used in this work, along with the version specifier and the corresponding VERITAS pipeline stage where they were used.

B VERITAS Stage Implementation Details

Stage 1: Review Discovery ClaimReviews are obtained via the Google Fact Check Tools API[2] and DataCommons[3]. Extracted review instances are refreshed by re-downloading ClaimReview data directly from the publisher websites. Language is determined using `langdetect`[4].

Stage 2: Publisher Identification We consider a fact-checking organization as credible if it is signatory of the International Fact-Checking Network (IFCN)[5] or member of the European Fact-Checking Standards Network (EFCSN)[6].

Stage 3: Article Scraping We scrape article content for each review using a dynamic engine based on Firecrawl[7] and Decodo[8]. Missing review publication and modification dates are obtained via HTML metadata.

Stage 4: Appearance Retrieval Appearances are scraped via social media APIs, `yt-dlp`[9], and Decodo.

Stage 6: Verdict Standardization To enable the ensemble LLMs to meaningfully return scoring values on the $[-1, 1]$ interval, we discretize it and assign semantically descriptive names to the values as follows:

- -1: **Negative** (certain)
- $-2/3$: **Negative** (rather certain)
- $-1/3$: **NEI** (negative, rather uncertain)
- 0: **NEI** (or fully uncertain)
- $1/3$: **NEI** (positive, rather uncertain)
- $2/3$: **Positive** (rather certain)
- 1: **Positive** (certain)

For media authenticity, if the decision maps to **Fabricated**, LLMs (and human annotators) may provide additional tags for finer-grained fabrication annotation, including:

- *AI-generated*: The media was (partially or entirely) synthesized by AI.
- *Manipulated*: The media is derived from a real recording, but was altered to change its meaning.
- *Forged*: The media is mostly or entirely a manually created invention.

When aggregating tags from multiple predictors, majority voting is applied for each tag individually (present vs. not present).

C VERITAS Composition and Subsampling

To compose the final release data, we subsample the available claims so that the number of **Intact** claims is equal to the number of **Compromised** claims for each quarter (holding true also for the longitudinal split). Moreover, text-only claims are rather overrepresented compared to media-based claims as the latter are dismissed more frequently for missing media. Therefore, we prioritize media-based claims over text-only claims during subsampling, setting 80% as the maximum share of media-based claims as determined by Dufour et al. (2024). To improve media quality, claims with media from original appearances are prioritized over claims with media from archives (typically saving media in reduced resolution) and claims with media from fact-checking articles (as the latter often contain annotations/edits).

[2] developers.google.com/fact-check/tools/api
[3] datacommons.org/factcheck
[4] pypi.org/project/langdetect
[5] ifcncodeofprinciples.poynter.org/signatories
[6] members.efcsn.com/signatories
[7] github.com/firecrawl/firecrawl
[8] decodo.com
[9] github.com/yt-dlp/yt-dlp

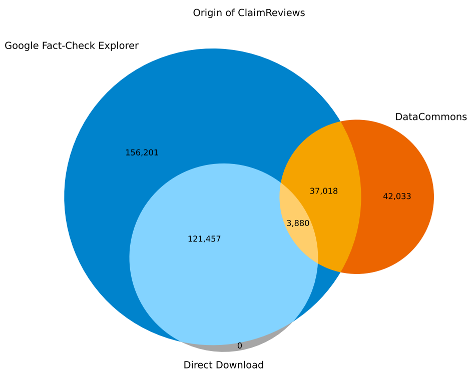


Figure 5: Origins of ClaimReviews

D VERITAS Statistics

D.1 ClaimReview

We obtained a total of $371,071\,K$ ClaimReviews starting from January 1, 2016. Three sources yielded the data: Google Fact-Check Explorer, Data Commons, and the fact-checking organizations themselves who added ClaimReview data to their published article webpages. The same ClaimReview can be obtained from multiple sources, where the ClaimReview directly downloaded from the publisher takes precedence over the others, as we expect it to be most up-to-date. Figure 5 shows the number of reviews returned by the three different sources. Figure 6 displays the number of reviews per quarter since the introduction of ClaimReview in 2016. Figure 7 shows the most common verdict labels as provided by the fact-checkers.

D.2 Appearances

We identified $434,947$ appearances. Figure 8 depicts the shares of appearances, containing the URL to the original source and/or to the archived record. In about 93% of the cases, we were able to identify the actual URL. Only for 7% of the appearances we could not resolve the original URL. Note that claims often have multiple appearances. Almost 40% of appearances also have a corresponding URL to an archived record. The top 10 platforms where appearances occur are listed in Fig. 9. The distribution of archiving services is shown in Fig. 10. The most prominent are Perma.cc (41.2%) and Archive.today (38.3%). The quarterly shares of appearances by platform are shown in Fig. 15b.

D.3 Claims

A total of $86,372\,K$ claims (original, rectified, and dismissed) were obtained by stage 5. The Figures 11 and 12 show the quarterly claim distributions for several different properties for the **released** $24\,K$ claims. The statistics for original (i.e., non-rectified and dismissed) claims are depicted in Figures 14 and 15.

D.4 Verdicts

Table 6 shows the total number of labels after discretizing scores into $[-1, -1/3)$ for negative, $[-1/3, 1/3]$ for neutral, and $(1/3, 1]$ for positive. Note that the totals across properties differ for two reasons: The number of media is different to the number of claims and the properties veracity and context coverage are evaluated only if the previous properties did not result in a Negative decision.

Property	Negative	NEI	Positive
Authenticity	3303	2439	5494
Contextualization	5405	431	5400
Veracity	6370	132	12095
Context Coverage	181	1	11945
Integrity	11953	94	11953

Table 6: Decision counts for all five properties.

E Metric Interpretation

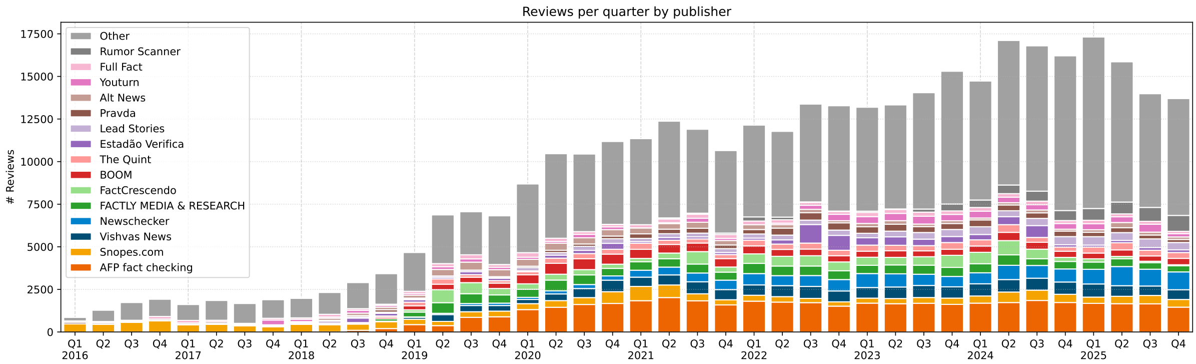
We use the Mean Squared Error (MSE) as the key performance indicator. Table 7 shows how to interpret the MSE values in the context of VERITAS.

F Baselines

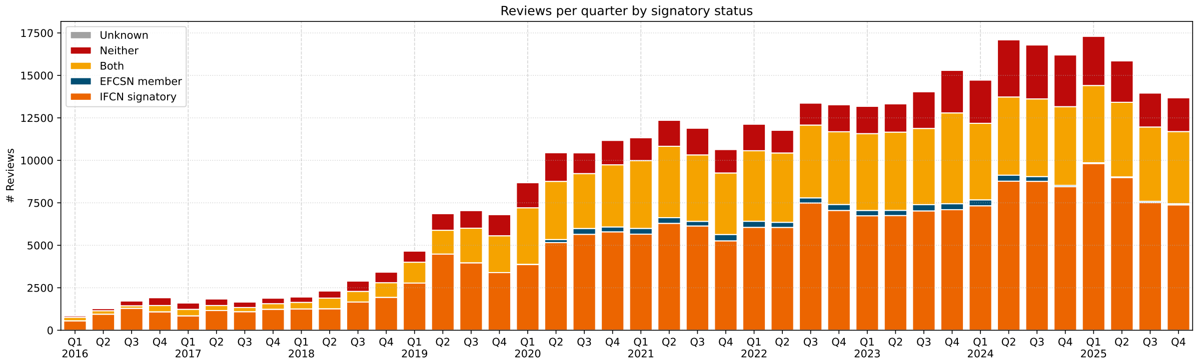
F.1 Web Search Tool

When provided with search access, models can query the internet via Google's Serper API, which returns the top 10 results per query. We apply two filters to retrieve results: temporal filtering (only content published before the claim date is retained) and domain blacklisting (fact-checking websites are excluded to prevent direct answer leakage, see supplementary material for the full list). For the top 3 filtered results, we attempt to scrape the full page content using Firecrawl; models receive both the search snippets from Serper and the complete scraped content when available. For each claim, up to 5 searches are possible.

The full results for the longitudinal split are shown in Table 8 and Figure 16. Figure 17 shows the results per modality.



(a) ClaimReviews by publishing organization.



(b) ClaimReviews by publisher signatory status.

Figure 6: ClaimReview statistics, as obtained by stages 1 and 2, showing *all* ClaimReviews exposed by Google and Data Commons retrieved by Dec 31, 2025.

MSE	Interpretation	Max. flipped	Equivalent to...
0.00	Perfect	0	Exact match.
0.04	Very Good	1 in 100	36 predictions being off by $1/3$ in 100 otherwise perfect predictions.
0.10	Good	1 in 40	Slightly better than being off by $1/3$ for all predictions.
1.00	Abstention	1 in 4	Being off by 1 always (roughly same as constantly predicting 0, i.e., ? **NEI**).

Table 7: How to interpret the Mean Squared Error (MSE) when evaluating with VERITAS. **Max. flipped** indicates, in a set of otherwise perfect predictions, the maximum number of flipped predictions, which are predictions that have a score difference of 2 to the target (i.e., complete opposite).

Figure 7: Raw ratings as provided by ClaimReview for English reviews.

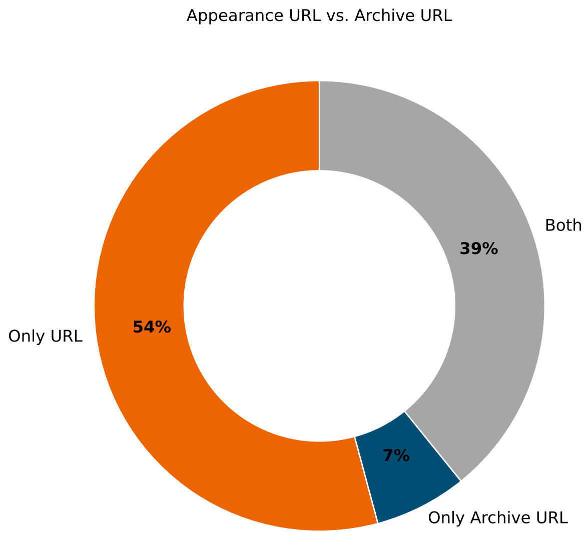


Figure 8: Appearances that have the original source URL, an archive URL, or both.

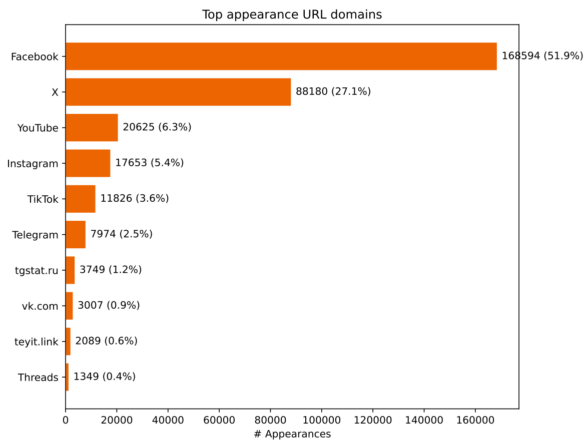


Figure 9: Appearances per platform, showing the top 10.

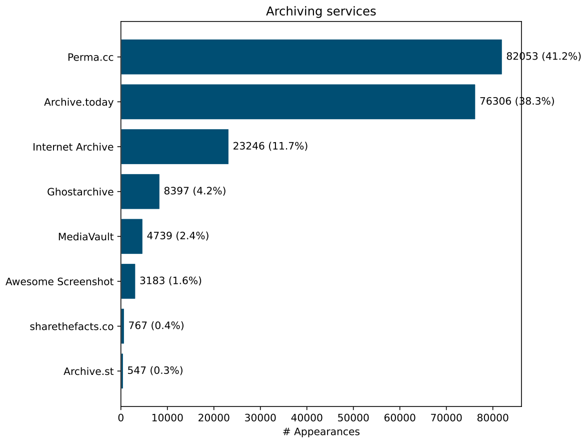
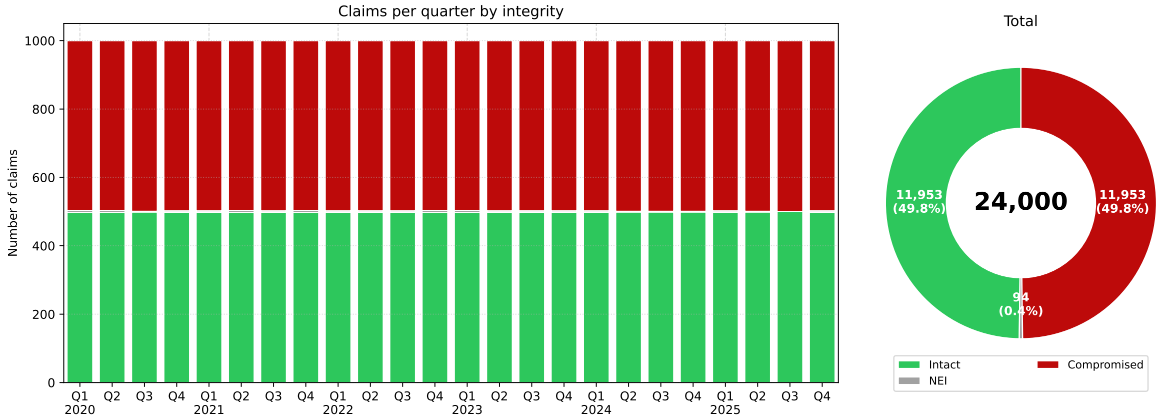


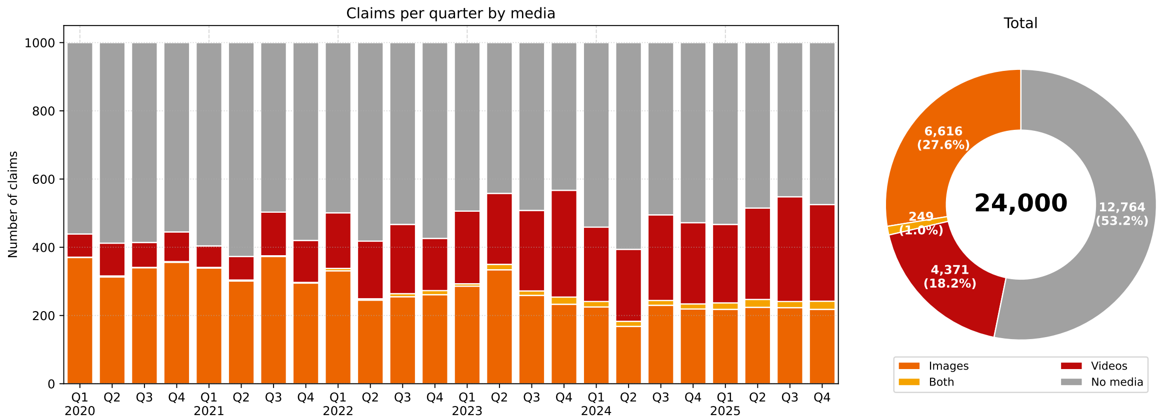
Figure 10: Appearances per archiving service.

Integrity	Search	**MSE (↓)**	**MAE (↓)**	**Acc. (↑)**
Gemini 2.0 Flash	-	0.63	0.59	46.72
Gemini 2.5 Flash	-	0.56	0.39	79.65
Gemini 3 Pro	-	**0.36**	**0.26**	**89.52**
GPT-4o	-	0.53	0.51	54.25
GPT-5.2	-	0.59	0.57	47.75
LLama 4 Maverick	-	0.81	0.61	55.63
Gemini 2.0 Flash	✓	0.60	0.51	59.13
Gemini 2.5 Flash	✓	0.50	0.41	71.73
Gemini 3 Pro	✓	**0.34**	**0.33**	**75.13**
GPT-4o	✓	0.54	0.47	61.83
GPT-5.2	✓	0.47	0.43	66.07
LLama 4 Maverick	✓	0.88	0.59	63.50
DEFAME (w/ GPT-5.2)	✓	0.50	0.46	61.61
LOKI (w/ GPT-5.2)	✓	0.88	0.59	62.01

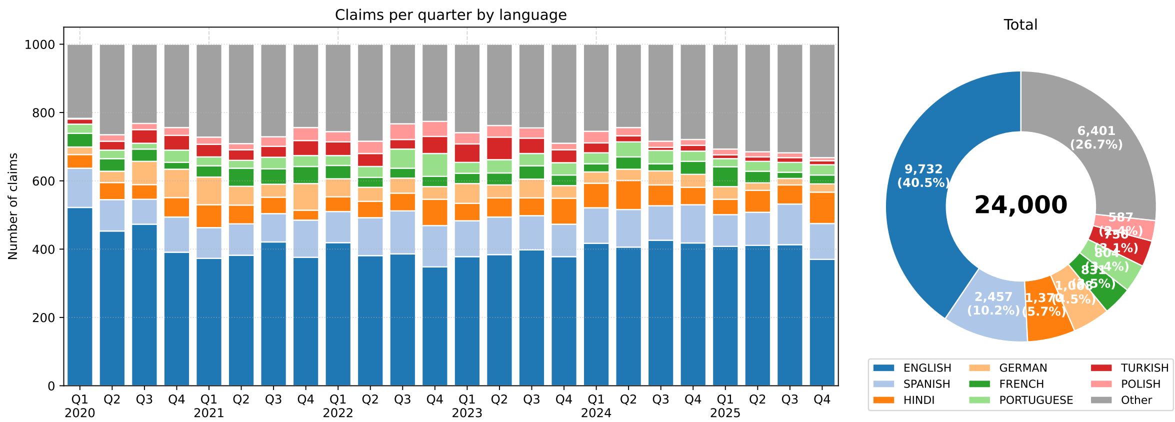
Table 8: Results on the **longitudinal data split**. Baselines have been tested using only parametric knowledge (without search tool) and using web search (with search tool). Additionally, DEFAME and Loki have been evaluated.



(a) Integrity shares.

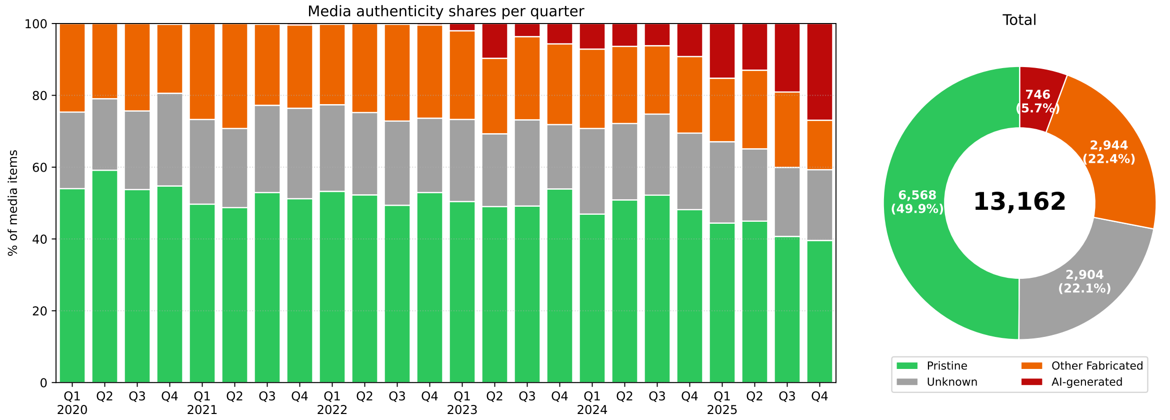


(b) Shares of claims featuring media.

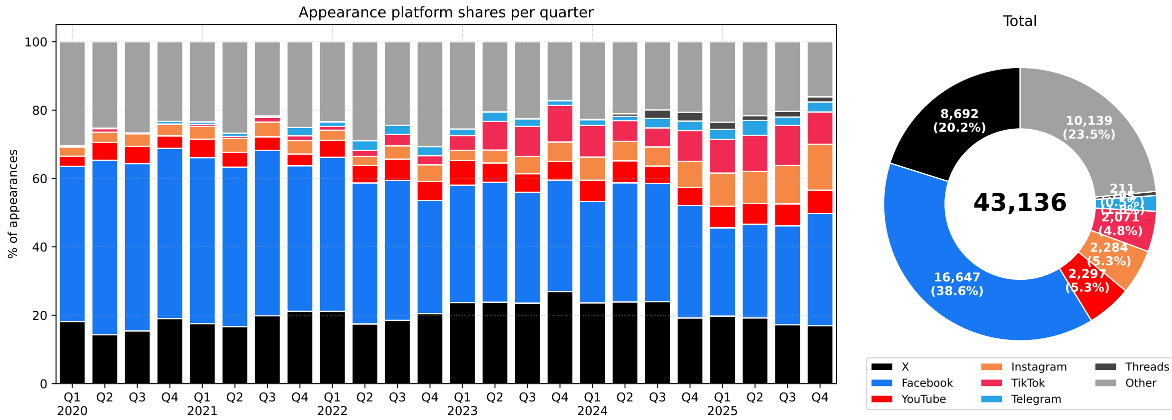


(c) Language shares.

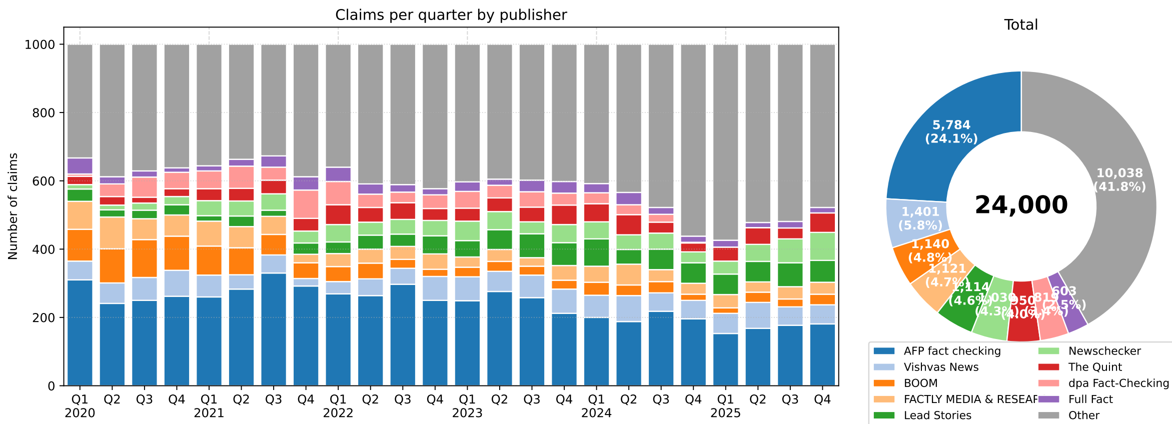
Figure 11: Statistics in the **final** VERITAS benchmark for all quarter splits.



(a) Media authenticity shares.



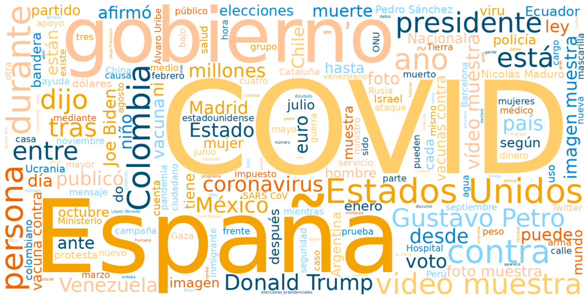
(b) Platform shares for all claim appearances.



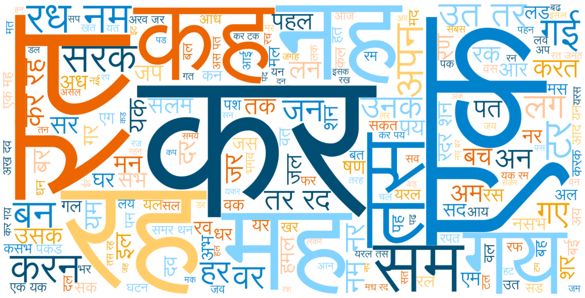
(c) Publisher shares.

Figure 12: Statistics in the **final** VERITAS benchmark for all quarter splits.

(a) English



(b) Spanish



(c) Hindi

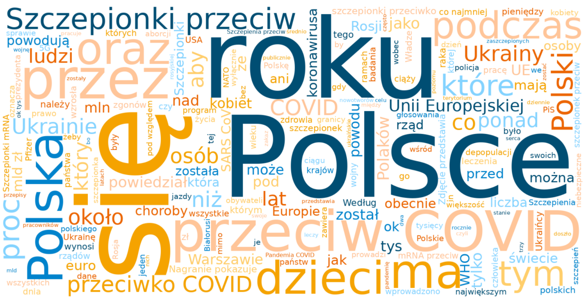
(d) German

(e) Portuguese

(f) French

(g) Turkish

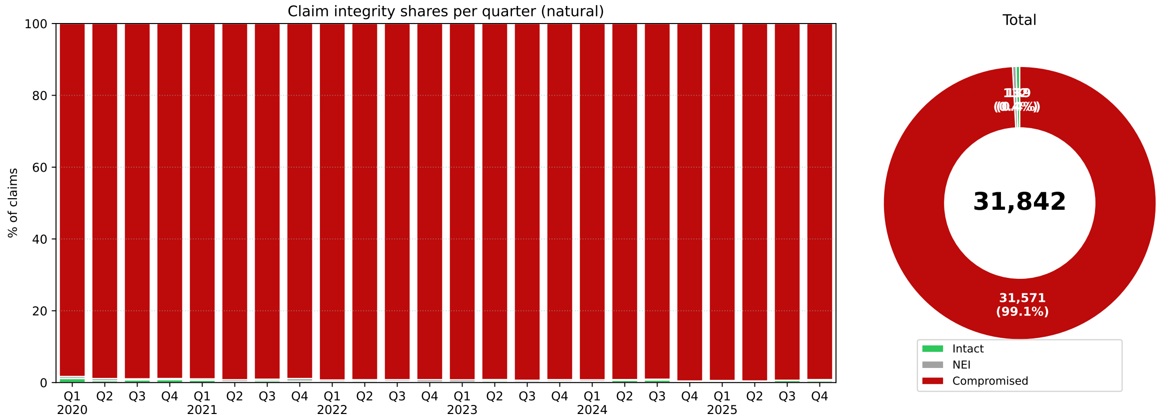
(h) Russian



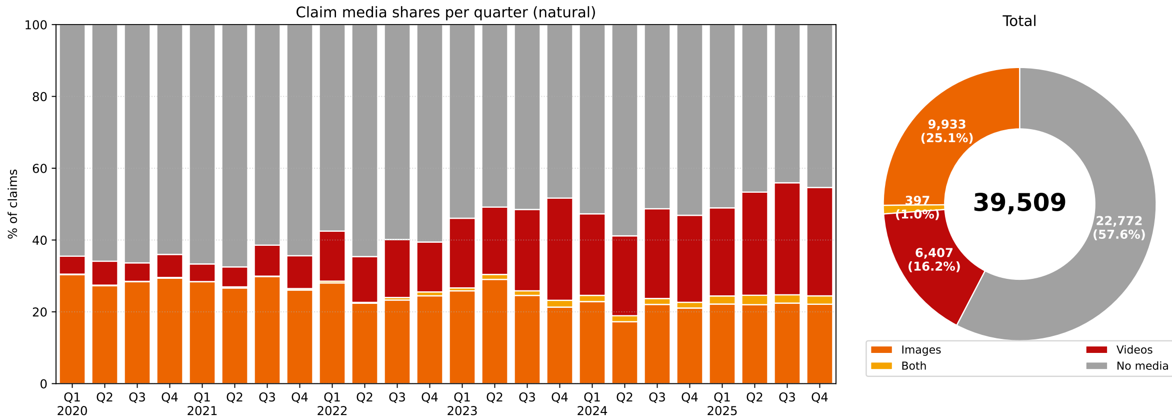
(i) Polish

(j) Arabic

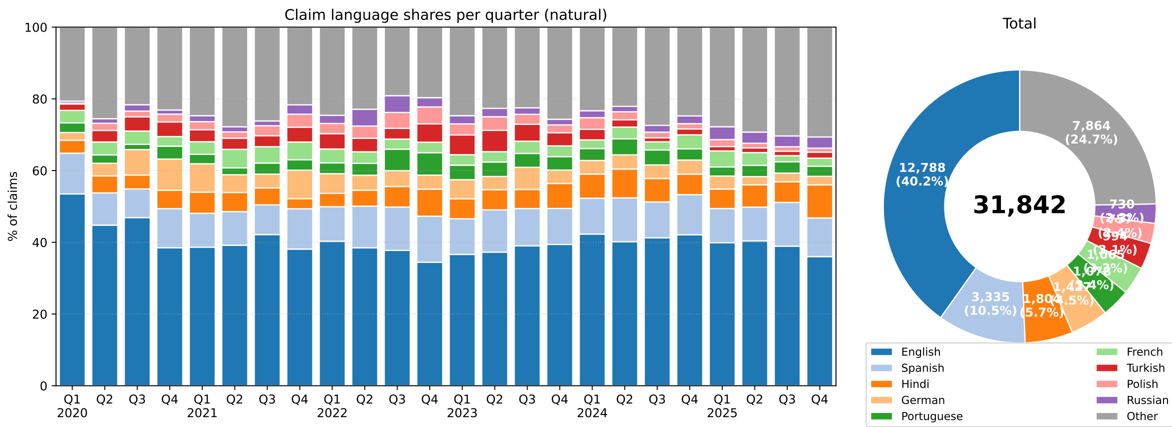
Figure 13: Common words occurring in VERITAS' claims for the ten most frequent languages.



(a) Integrity shares.

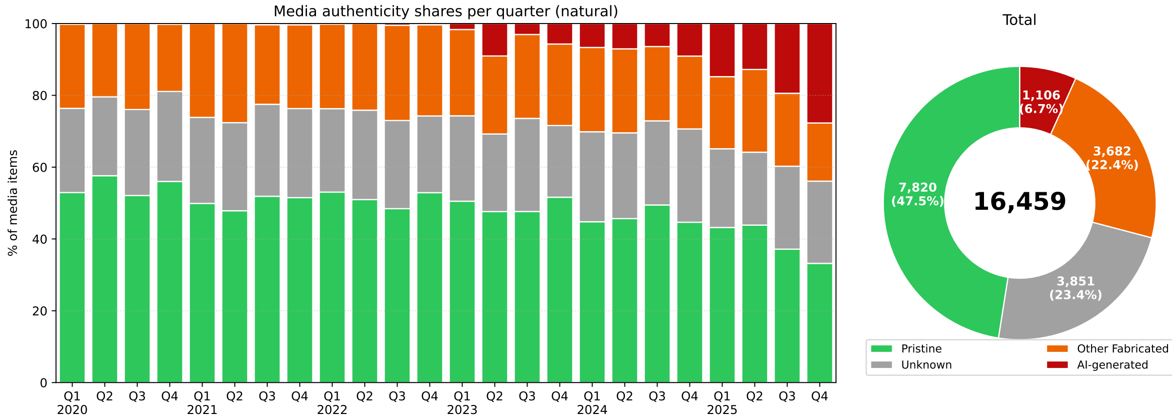


(b) Shares of claims featuring media.

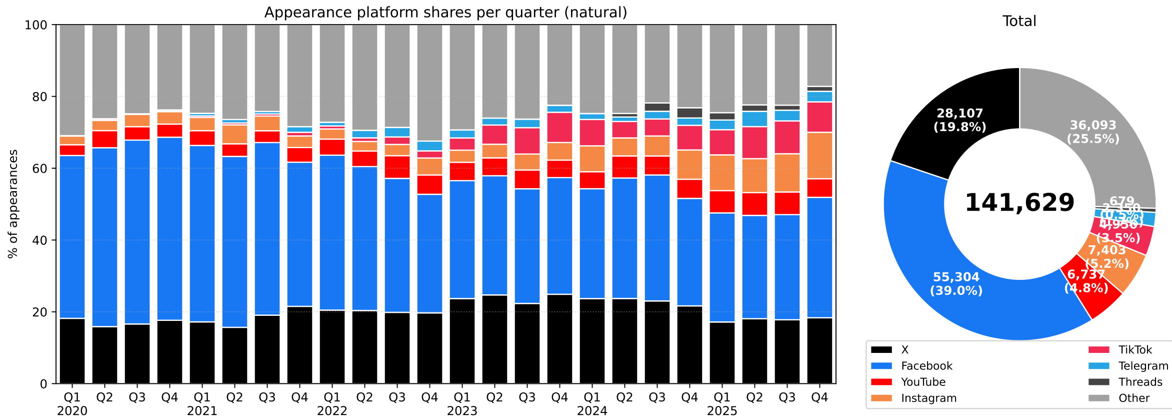


(c) Language shares.

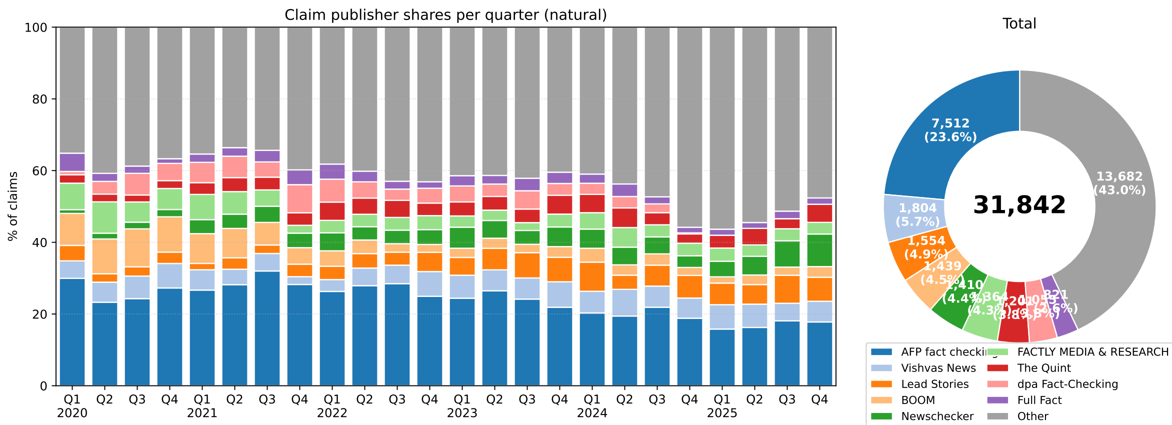
Figure 14: **Natural** data statistics as obtained before stage 7, i.e., before balancing and sampling.



(a) Media authenticity shares.

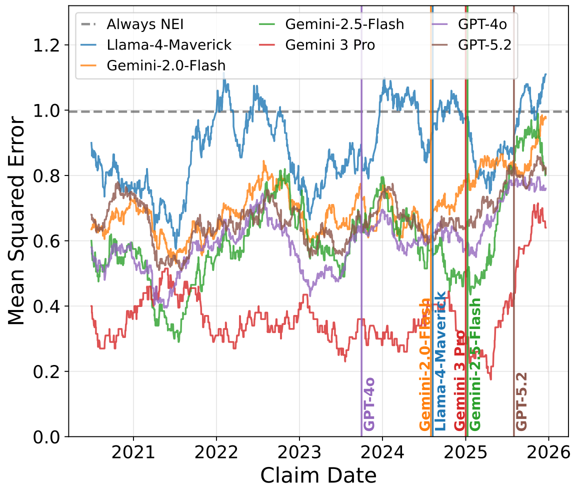


(b) Platform shares for all claim appearances.

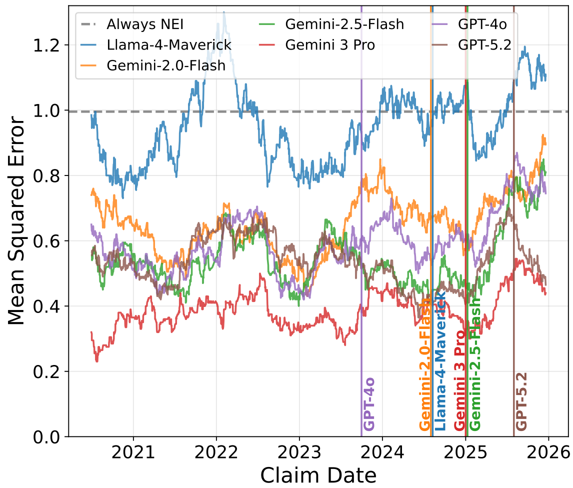


(c) Publisher shares.

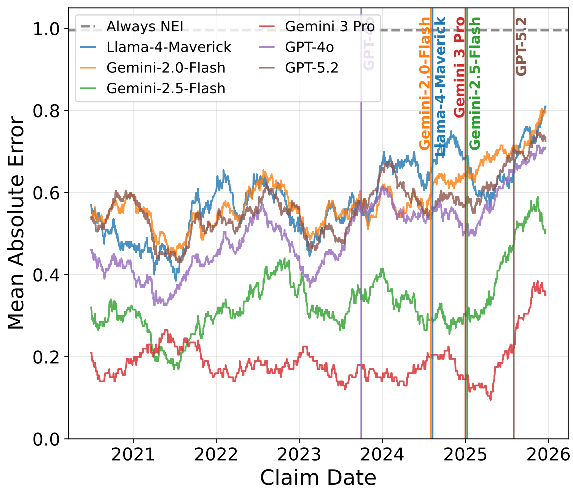
Figure 15: **Natural** data statistics as obtained before stage 7, i.e., before balancing and sampling.



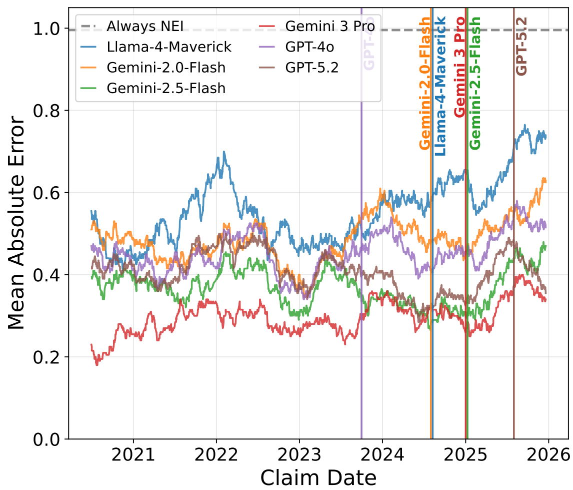
(a) MSE **without** web search.



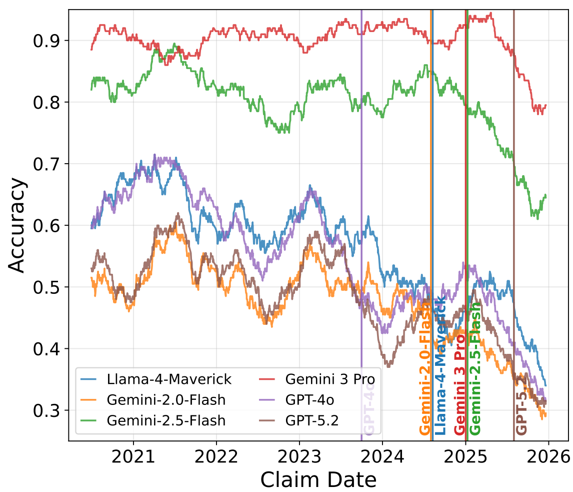
(b) MSE **with** web search.



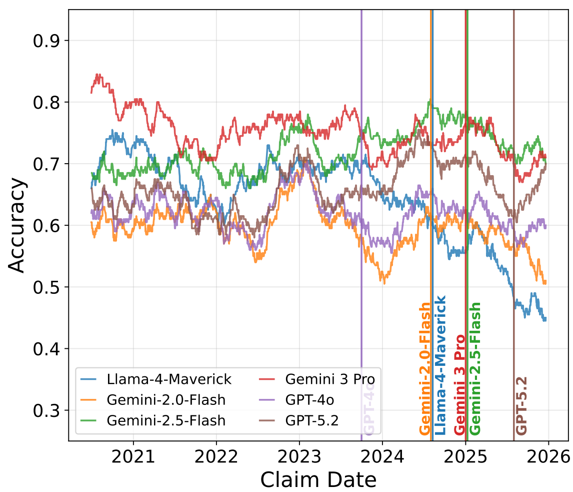
(c) MAE **without** web search.



(d) MAE **with** web search.

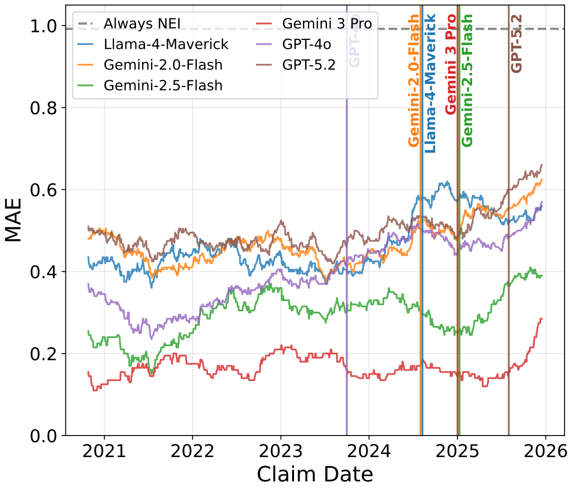


(e) Accuracy **without** web search.

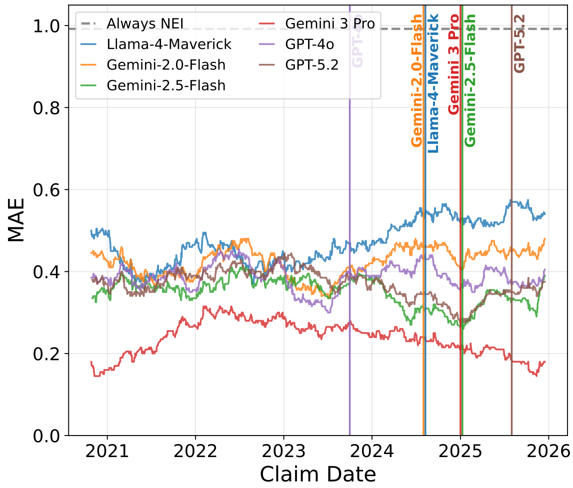


(f) Accuracy **with** web search.

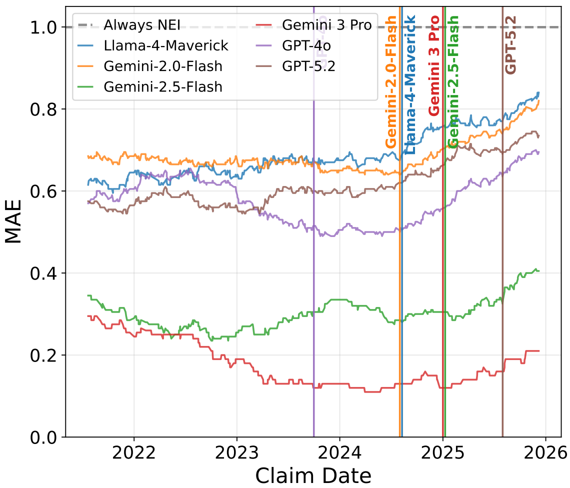
Figure 16: Baseline results on the **longitudinal split** for all three metrics Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy (by 3-bin discretization). All plots use a 200-claim moving average window. Vertical lines indicate knowledge cutoff dates.



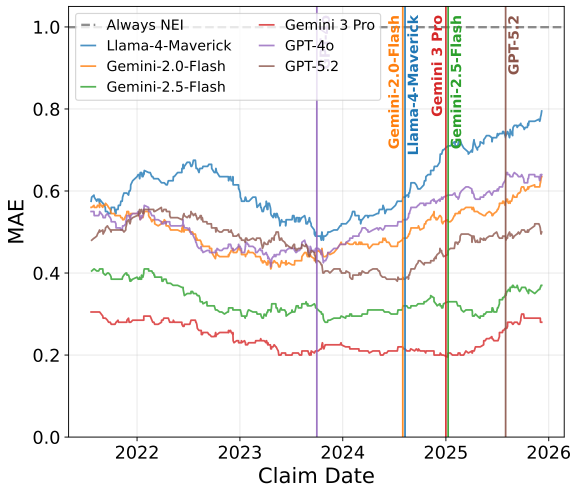
(a) MAE for text-only claims **without** web search.



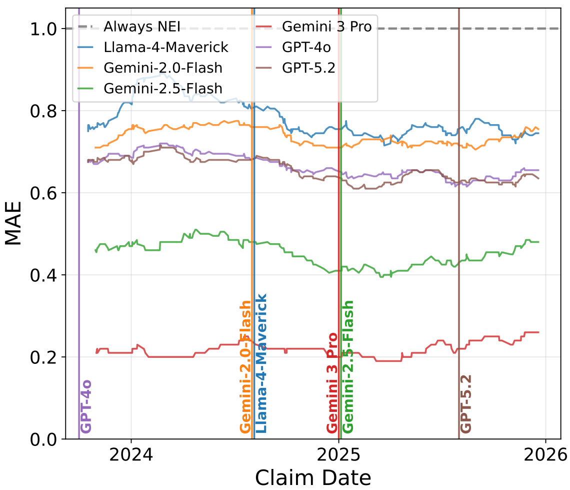
(b) MAE for text-only claims **with** web search.



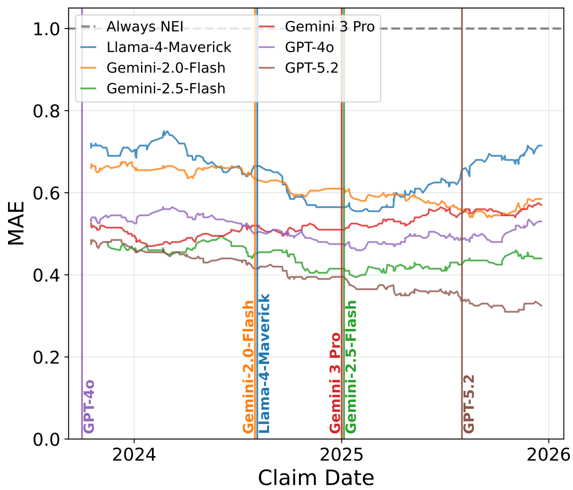
(c) MAE for text+image claims **without** web search.



(d) MAE for text+image claims **with** web search.



(e) MAE for text+video claims **without** web search.



(f) MAE for text+video claims **with** web search.

Figure 17: Baseline results on the **longitudinal split** for modality specific Mean Absolute Error (MAE). All plots use a 200-claim moving average window. Vertical lines indicate knowledge cutoff dates.

G Human Evaluation

G.1 Additional Setup Details

Annotation Organization. We selected 12 annotators having at least graduate-level education in Computer Science. 372 annotations (verdict assessment per claim) were gathered, spanning 9 different languages. Annotation was done without fees, although two small presents were raffled among the annotators to honor their contribution. The annotation procedure was introduced in a 14 min annotation tutorial video[10]. We dismissed annotations with score difference larger than 1, same for claims that received less than 2 annotations.

Annotation Process. Annotators first read the original fact-checking article associated with a claim and perform a manual validation, assessing that the claim (i) is unambiguous, (ii) does not contain text exposing or hinting at the verdict, (iii) any attached media does not contain overlays or labels from the fact-checking article, and (iv) all media referenced in the claim text is attached. Claims failing any criterion are discarded with the failed checks recorded.

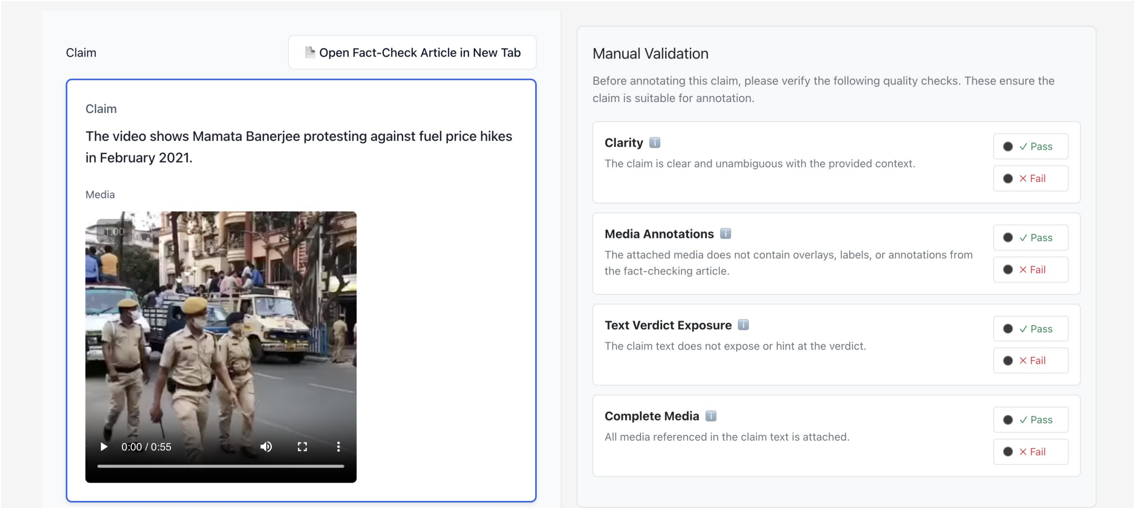
The evaluation then proceeds sequentially through the properties, with the flow conditioned on intermediate results. If the claim contains media, annotators first evaluate each media item's authenticity and contextualization. Veracity is assessed only if no media exists or all media contextualizations score above the negative threshold ($\geq -1/3$). Context coverage is assessed only if veracity is evaluated and judged as true. This conditional structure reflects that a negative decision on contextualization or veracity already compromises the claim's integrity, making subsequent assessments redundant. See Fig. 18 for a full annotation example.

For each property, annotators provide a judgment on a seven-point scale with three confidence levels (certain, rather certain, rather uncertain) for each direction, plus a neutral midpoint denoting total uncertainty. Each judgment requires a written explanation. When annotators select a confidence level below "rather certain," they must additionally specify the source of their uncertainty from predefined options: lack of evidence in the article, contradicting evidence, or other (free-text).

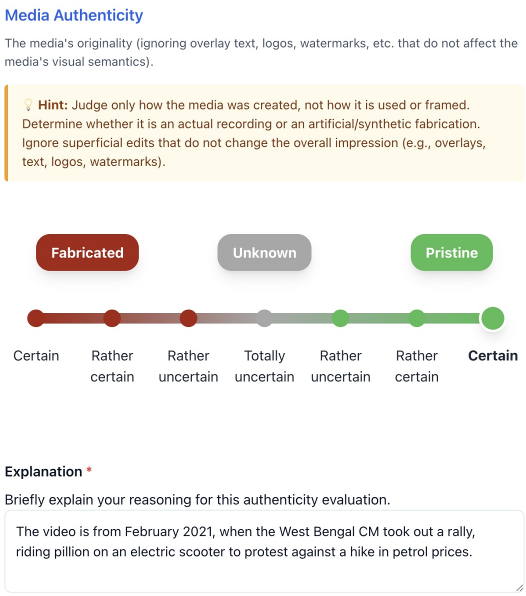
[10]Excluded for anonymity reasons.

G.2 Additional Results

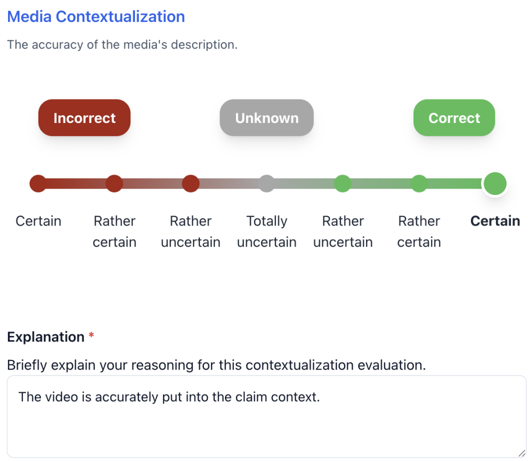
Figure 19 compares human annotation with automated annotation for all 5 properties, including an aggregated view summing up all properties.



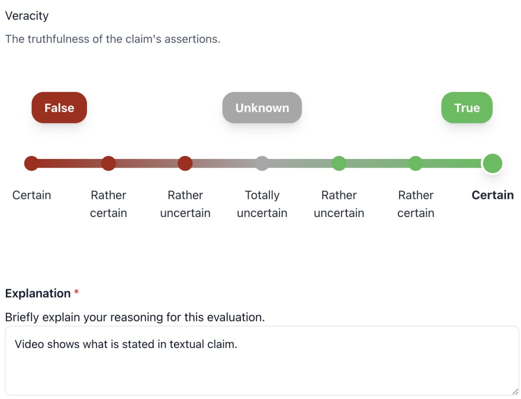
(a) Step 1: Validate that claim fulfills all quality requirements.



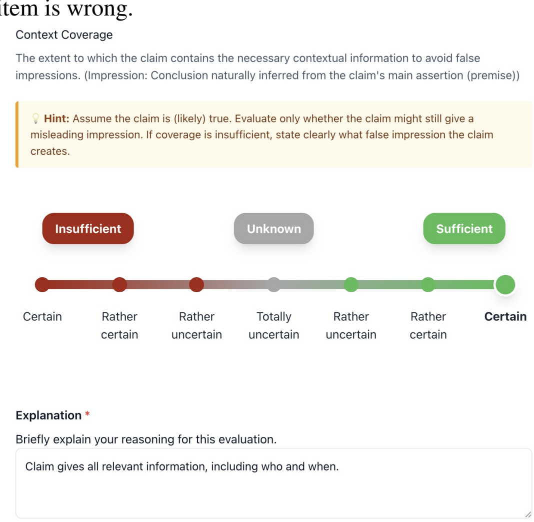
(b) Step 2: Assess media authenticity.



(c) Step 3: Assess media contextualization. Terminate annotation if the contextualization of at least one media item is wrong.



(d) Step 4: Assess claim veracity. Terminate annotation if the veracity is false.



(e) Step 5: Assess context coverage.

Figure 18: Overview of the human annotation process: After validating quality requirements, annotators assess all properties sequentially with early termination. (b)-(e) only show the right column of the annotation view, the left column always shows the claim and metadata like in (a).

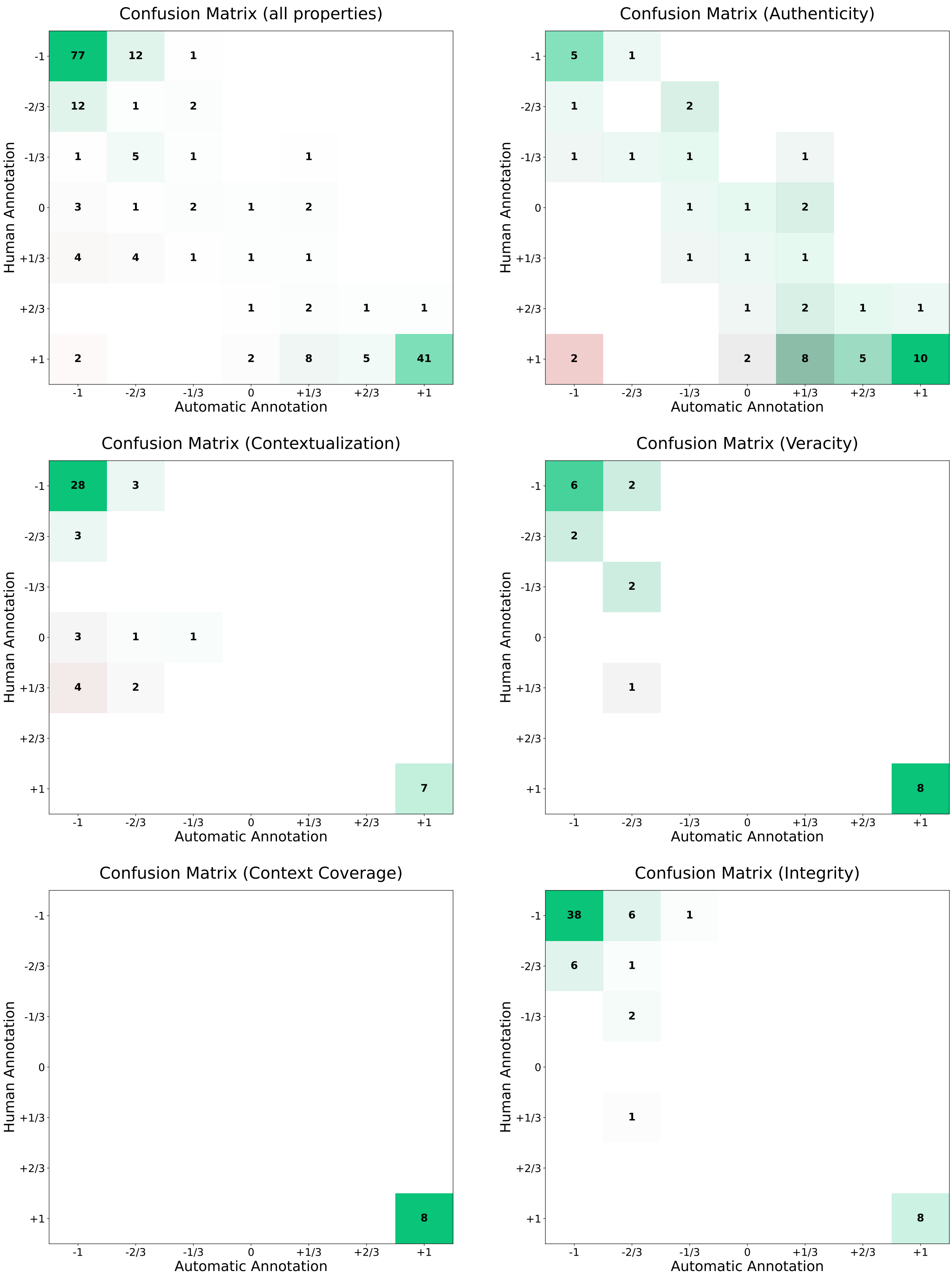


Figure 19: Confusion matrices comparing human annotations with VERITAS' automated annotations.