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

Mark Rothermel **Marcus Kornmann** **Marcus Rohrbach** **Anna Rohrbach**
Multimodal AI Lab
Technical University of Darmstadt
hessian.AI
Correspondence: mark.rothermel@tu-darmstadt.de

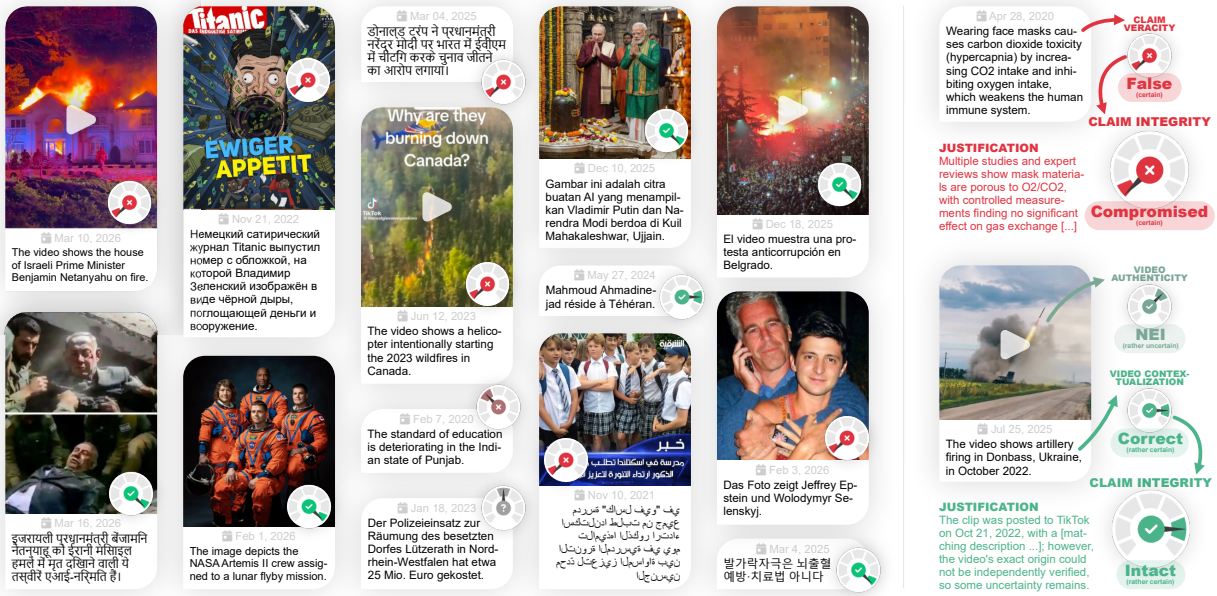


Figure 1: The VERITAS benchmark: Example claims including media, claim date, and claim integrity score. The two claims on the right showcase lower-level annotations of media/claim properties used to infer the overall integrity. Each annotation contains a justification. New claims are added quarterly 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 25,000 real-world claims from 104 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. The code and data are publicly available under veritas.mai.informatik.tu-darmstadt.de.

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

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42066–42100

higher engagement and spreads faster, particularly in the context of misinformation (Li and Xie, 2020; Zannettou et al., 2018; Wang et al., 2021). 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. Manual 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 Fig. 5, demonstrating a performance drop for claims published after the 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. Fig. 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 an evaluation metric to reward near-correct predictions and strongly penalize flipped decisions.

Finally, we benchmark state-of-the-art multimodal AFC systems and strong LLM baselines, revealing substantial room for improvement. We made the benchmark data and pipeline code available at veritas.mai.informatik.tu-darmstadt.de and commit to extending VERITAS in each subsequent 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 as AVERITEC (Schlichtkrull et al., 2023) address-

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
MR² (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
DGM⁴ (Shao et al., 2023)	MD	✓	-	-	✓	-	✓	1	News	-	-	230 K	2 Classes
VeriTaS (Ours)	CV	✓	✓	✓	✓	✓	✓	54+	✓	✓	✓	25 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.

ing 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 either 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, among 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

8 months ago. Fatahi Bayat et al. (2025) propose a dynamic, text-only benchmark that consists not of claims but of prompts of varying complexity 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 dynamic, multimodal, and multilingual benchmark for evaluating Multimodal Automated Fact-Checking (MAFC) systems. It gets extended with recent, real-world claims each quarter, combining text, images, and videos. The data is gathered 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 (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 and model availability (see App. A for the used LLM versions). 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 code release for the prompts. Fig. 2 summarizes the seven autonomous 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 automatically 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 398 K reviews published between Jan 1, 2016 and Mar 31, 2026, 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 848 distinct publishers via review URLs, and retain reviews only if they originate from credible fact-checkers (cf. App. B), dismissing about 64 K (16%) of the reviews not meeting this criterion. **Output.** 335 K 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 8.8 K (4.2%) inaccessible articles as well as 7.3 K (3.5%) unrealistically short articles. **Output.** 208 K reviews[2] 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. I.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. 105 K (52.8%) Reviews without any valid appearance are discarded. **Output.** 94 K reviews[2] 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 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 > 0.85 in the embedding space of the CLIP variant `Qdrant/clip-ViT-B-32-vision`. Videos are inspected with GEMINI 2.5 FLASH, images with GPT-5 MINI, to exclude irrelevant content. Media precedence follows: original appearance $>$ archived appearance $>$ article media. At most 4 media 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 72 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. 480 (0.1%) 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.** 72 K validated, well-formed claims[2] with associated media, if relevant.

[2]At the time of writing, there remain unprocessed reviews at this stage since the gathered data was already sufficient.

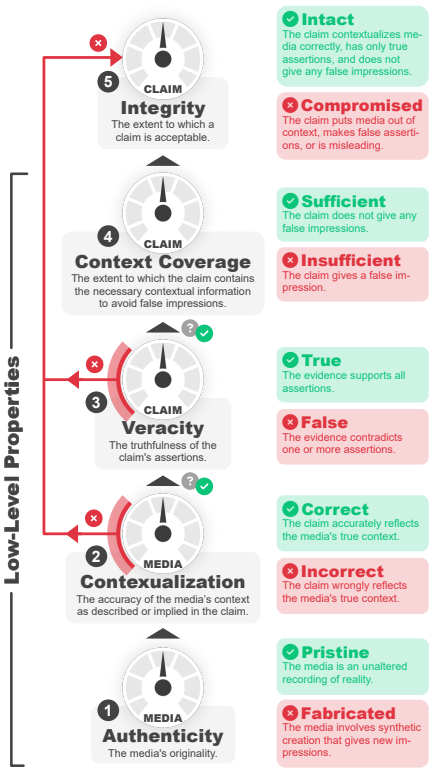


Figure 3: Verdict derivation (Stage 6), assessing properties (1) to (4). Negative decisions in (2) or (3) result in early termination. Definitions are shown on the right.

3.6 Stage 6: Verdict Standardization

Goal. Standardize heterogeneous fact-check ratings. Fact-checking organizations employ heterogeneous rating schemes that often conflate multiple dimensions, using labels such as "half true," "legend," or "three Pinocchios" (see Fig. 10). 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 known 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-scoring property among (2–4), 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 318 (0.9%) claims that receive an ensemble prediction with an internal score difference exceeding 1 to keep only instances with high inter-model agreement. Additionally, each ensemble member provides a one-paragraph justification for its 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.** 36 K claims[2] 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. 17a. Only $\sim$ 0.3% of ClaimReviews yield ✅ Intact claims. Thus, we generate about 32 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 7.3 K (23.1%) rectified claims for exclusion. Finally, rectified claims are reevaluated using Stage 6. Claims with integrity $> 1/3$ are kept and replace their originals; all 178 (0.6%) others are discarded along with 3.1 K (9.9%) claims with insufficient ensemble agreement. **Output.** 17 K consistent, shareable, intact rectified claims.

4 Results

4.1 VERITAS Statistics

VERITAS currently contains 25 K claims spanning 25 quarterly splits from Q1 2020 to Q1 2026. Each quarter comprises 1 K claims, balanced so that the number of ✅ Intact and ❌ Compromised instances is equal. Earlier periods are excluded due to insufficient review coverage. The most recent quarter is generally the only one suitable for reliable benchmarking. In addition to the quarterly splits, we release a budget-friendly *longitudinal split* containing 2,500 balanced claims (100 per quarter), enabling temporal analysis across the full time span. Summary statistics are reported in Fig. 4 and App. I.4.

VERITAS contains 8,692 images and 5,334 videos while it covers 54 languages, with 37 occurring at least 50 times. English is most frequent (39.0%), followed by Spanish (10.9%) and Hindi (5.8%), see Fig. 14c for more languages. Claim appearances are predominantly sourced from Facebook (38.4%) and X/Twitter (20.5%), while all other platforms contribute less than 6% each. AFP Fact Checking accounts for the largest share of reviewed claims (24.8%), with all other organizations contributing below 6% each.

4.2 Validation through Human Evaluation

We validate the outputs from stages 6 and 7 (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 or seven bins, respectively. MSE is our primary metric, as it penalizes severe confusions more strongly (App. D). MSE ≤ 0.04 indicates very high agreement (App. D). Results are

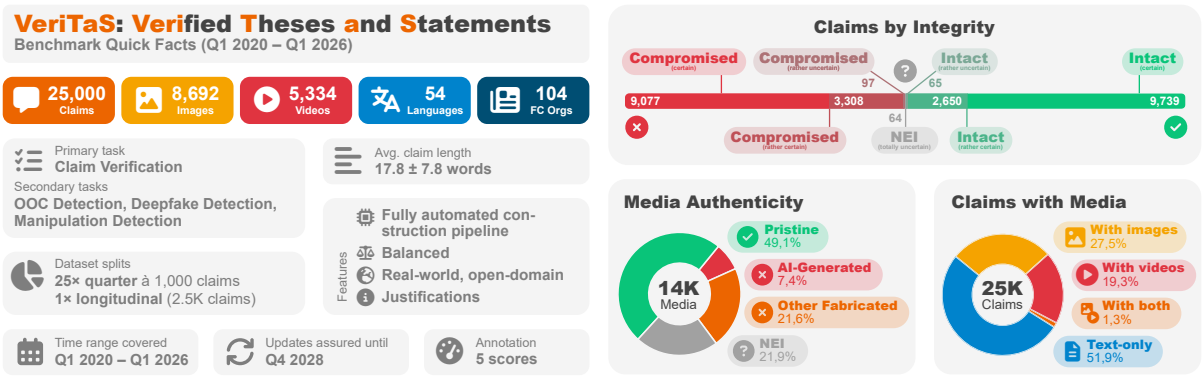


Figure 4: VERITAS dataset statistics, see also Tab. 8.

Integrity	N	Error Rates (↓) MSE	MAE	Accuracy (↑) 7-bin	3-bin
VERITAS	204	**0.034**	0.102	69.1	**97.5**
- w/ ensemble, w/o filtering	207	**0.035**	0.105	68.6	**97.6**
- GPT-5.2 alone	207	0.076	0.184	51.2	95.7
- GEMINI 3.1 PRO alone	207	0.071	0.099	**73.4**	95.7
- CLAUDE SONNET 4 alone	207	0.048	**0.091**	72.5	97.1
- LLAMA 4 MAVERICK alone	207	0.042	0.103	66.7	97.1

Table 2: Human evaluation results over N claims. The scores reflect agreement between human annotation and automated VERITAS annotations for integrity for six different settings. **MSE** = Mean Squared Error (the main metric), **MAE** = Mean Absolute Error, **Accuracy** = share of matches of n-bin discretized scores (in %).

summarized in Tab. 2, where we also include five ablated VERITAS variants: *w/ ensemble, w/o filtering* (keeping claims with high ensemble disagreement) and just one of the four ensemble members as the sole predictor. The evaluation shows a very strong agreement with an MSE below 0.04, corresponding to roughly one flipped prediction among 100 otherwise correct judgments. The observations are confirmed by confusion matrices (Fig. 25) and the results for low-level properties (Tab. 10).

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. 2 indicate the benefits of the ensemble approach and the agreement filtering. Moreover, evaluators were asked to flag problematic claims. Only a small minority ($\sim$ 5.1%) exhibited quality concerns (primarily missing media, otherwise clarity issues), indicating that the normalized/rectified claims are generally well-formed and realistic. See App. K for more details on the evaluation.

4.3 Benchmarking AFC Methods on VERITAS latest quarter

We analyze baselines and current AFC methods on the Q1 2026 split to assess performance on the most recent data—particularly after the knowledge cutoff of all tested methods. We consider various foundation models: CLAUDE OPUS 4.6, GPT-5.2 and GEMINI 3.1 PRO as state-of-the-art multimodal models; GPT-4O as an earlier-generation model to study knowledge-cutoff effects; GEMINI 3 FLASH as a mid-sized LLM, as well as LLAMA 4 MAVERICK, GEMMA 4 (31B), and QWEN 3.5 (397B) as representative open-source models. Each model is evaluated once using parametric knowledge only and once with web search augmentation, see App. J.1 for the tool implementation details. Note that *all* evaluations on VERITAS require excluding evidence sources published after the claim's release date to avoid temporal leakage (Glockner et al., 2022).

GEMINI 3.1 PRO and GEMINI 3 FLASH process videos natively. For all other models, we represent videos via five evenly spaced frames and a speech transcript. Additionally, we compare against two recent fact-checking systems, DEFAME (Braun et al., 2025) and LOKI (Li et al., 2025b), both with three different backbones in Tab. 3.

Across all baselines, CLAUDE OPUS 4.6 achieves the strongest MSE in both evaluation settings, even surpassing the AFC-specialized models DEFAME and LOKI. The second-best model, GEMINI 3 FLASH, even surpassing GEMINI 3.1 PRO, trails by +0.18 and +0.09 MSE, depending on retrieval access. Open-source LLAMA 4 MAVERICK performs the worst among LLMs, also under search augmentation. In contrast, open-source GEMMA 4 (31B) and QWEN 3.5 (397B)both perform on par or better

Method	Search	Overall Results				MSE on Subsets (↓)						
		Errors (↓)		Accuracy (↑)		By Media			By Language		By Integrity	
		MSE	MAE	7-bin	3-bin	w/ images	w/ videos	text-only	English	non-English	Intact	Compr.
GEMINI 3 FLASH	-	0.632	**0.409**	**60.3**	**81.4**	0.553	0.743	0.625	0.604	0.648	1.097	0.167
GEMINI 3.1 PRO	-	0.636	0.412	58.9	81.1	0.508	0.751	0.659	0.617	0.647	1.210	**0.049**
CLAUDE OPUS 4.6	-	**0.453**	0.471	30.1	62.2	**0.451**	**0.509**	**0.361**	**0.450**	**0.455**	**0.572**	0.337
GPT-4O	-	0.826	0.690	25.4	45.3	1.050	0.759	0.560	0.967	0.760	1.418	0.247
GPT-5.2	-	0.701	0.691	12.8	42.0	0.644	0.820	0.571	0.678	0.710	1.179	0.240
LLAMA 4 MAVERICK	-	0.943	0.767	18.1	44.2	1.127	0.949	0.583	0.966	0.931	1.270	0.635
GEMMA 4 (31B)	-	0.675	0.465	51.5	77.2	0.590	0.791	0.703	0.632	0.697	1.059	0.290
QWEN 3.5 (397B)	-	0.887	0.563	51.3	69.8	0.828	1.070	0.587	0.911	0.881	1.404	0.373
GEMINI 3 FLASH	✓	0.275	0.237	**67.2**	**90.5**	0.250	0.332	**0.158**	0.268	0.279	0.371	0.173
GEMINI 3.1 PRO	✓	0.388	0.316	61.1	82.9	0.335	0.443	0.277	0.397	0.386	0.643	**0.132**
CLAUDE OPUS 4.6	✓	**0.183**	**0.221**	60.8	89.6	**0.134**	**0.210**	0.196	**0.172**	**0.188**	**0.161**	0.202
GPT-4O	✓	0.841	0.615	38.4	59.6	1.141	0.702	0.547	0.993	0.770	1.362	0.333
GPT-5.2	✓	0.353	0.422	22.8	77.1	0.271	0.437	0.313	0.261	0.393	0.544	0.166
LLAMA 4 MAVERICK	✓	0.863	0.712	22.0	50.2	0.934	0.904	0.535	0.811	0.878	1.032	0.707
GEMMA 4 (31B)	✓	0.360	0.295	60.7	87.2	0.328	0.429	0.236	0.331	0.374	0.277	0.438
QWEN 3.5 (397B)	✓	0.318	0.296	57.8	86.7	0.263	0.354	0.254	0.237	0.354	0.393	0.240
DEFAME W/ GEMINI 3 FLASH	✓	0.450	0.320	**62.2**	87.2	0.318	0.589	0.319	0.387	0.472	0.507	0.391
DEFAME W/ CLAUDE OPUS 4.6	✓	**0.282**	**0.277**	58.1	**88.3**	**0.157**	**0.411**	**0.283**	**0.213**	**0.314**	**0.440**	0.123
DEFAME W/ GEMMA 4 (31B)	✓	0.465	0.349	59.3	83.6	0.344	0.599	0.316	0.329	0.527	0.491	0.441
LOKI W/ GEMINI 3 FLASH	✓	1.487	0.852	39.4	54.8	1.422	1.541	1.659	1.295	1.561	2.821	0.151
LOKI W/ CLAUDE OPUS 4.6	✓	1.638	0.920	38.0	52.0	1.648	1.797	1.359	1.657	1.623	3.195	**0.080**
LOKI W/ GEMMA 4 (31B)	✓	1.109	0.800	23.0	42.5	1.025	1.214	0.894	0.871	1.215	0.892	1.336

Table 3: Model evaluation on the **Q1 2026 split** with(out) search tool, single runs, with **best** and second best.

than top-tier GEMINI 3.1 PRO. Surprisingly, despite its specialization, LOKI underperforms all other models. Subset results reveal several observations: (1) Video claims are the most difficult to fact-check for almost all models, (2) only GPT-5.2 and QWEN 3.5 (397B) seem to noticeably struggle with non-English claims, and (3) many models exhibit a strong bias towards rating claims as ✗ Compromised.

4.4 The Role of Knowledge Cutoff

We evaluate the same models on the longitudinal VERITAS split. Fig. 5 shows the results. All evaluated models exhibit a marked increase in MSE after their respective Knowledge Cutoff Dates (KCDs), as one can see for the consistently increasing solid curves. 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 3.1 PRO, whose MSE increases from around 0.3 to over 0.7. We can rule out the increasing share of videos as a confounding factor, since the same phenomenon is observed for text-only claims, see Fig. 22 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

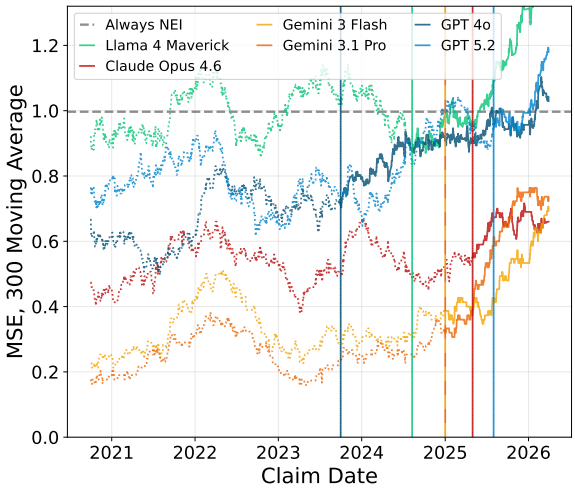


Figure 5: Baseline MSE performance without web search on the **longitudinal split**, single runs, smoothed with moving average by claim date. Lower is better. Vertical lines indicate knowledge cutoff dates, solid curves post-cutoff performance.

over time, exceeding 25% of claims by Q4 2025, while the proportion of pristine media correspondingly declines (Fig. 18). The fraction of claims containing media—particularly videos—also increases over time (Fig. 17). Platform coverage shifts are observable, with a declining share of Facebook and a growing share of Instagram (Fig. 18). Finally, the volume of all published ClaimReviews rises until Q1 2025 and decreases thereafter (Fig. 9).

5 Discussion

Data leakage severely undermines static AFC benchmarks. The increase in error rate of models after their Knowledge Cutoff Date (KCD) provides strong empirical evidence that LLMs partially encode real-world claims and verdicts from before their KCD. Vice versa, benchmarks dominated by pre-KCD data systematically overestimate model performance. This "*pre-cutoff inflation*" phenomenon is concerning since realistic fact-checking usually targets newly emerging claims. 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.

Nature and Validity of Rectified Claims. High-quality AFC benchmarks should be grounded in real claims, as purely synthetic data fails to capture the diversity and realism of misinformation and raises ethical concerns. In contrast, VERITAS generates ✅ Intact claims conditioned on *expert ground truth* through rectification. Although roughly 50% of claims are generated this way, they are fundamentally different from synthetic claims in prior work. VERITAS' rectified claims are directly grounded in real-world fact-checking articles and their associated evidence, rather than being artificially constructed or designed to mislead. Moreover, they undergo a multi-stage validation process, including filtering, where a substantial fraction is discarded, as well as an additional *Integrity* check. Finally, human evaluation indicates that only a small minority of all claims ($\sim 5.1\%$) exhibit quality concerns. Taken together, these results suggest that rectified claims are not synthetic artifacts in the conventional sense, but rather evidence-grounded corrections of real-world claims that preserve realism while enabling dataset balancing.

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 (CLAUDE OPUS 4.6, MSE $= 0.18$) corresponds to approximately one flipped prediction in 22 otherwise perfect ratings. Specialized AFC systems appear to degrade the performance of the tested LLMs when used as backbones. These results underscore that multimodal AFC remains far from solved. Finally, the fact that GEMMA 4 (31B) performs on par with substantially larger proprietary models such as GPT-5.2 and GEMINI 3.1 PRO suggests that model scale alone is not required to achieve decent performance.

Reliance on Proprietary LLMs. Dependence on proprietary models raises concerns regarding reproducibility. To assess sensitivity to the underlying backbone, we partially regenerated a subset of Q4 2025 using a substantially different LLM family (see App. H). The resulting claims preserve the same core assertions with only minor wording differences, and ensemble-based verdicts remain highly consistent. While this does not yet fully demonstrate reproducibility with open-weight models, these findings suggest that VERITAS is largely invariant to the specific proprietary backbone.

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 $25\,\mathrm{K}$ claims toward the full set of roughly $400\,\mathrm{K}$ reviews would further improve coverage and analytical power. Since claims can be ambiguous, future work should find strategies to address multiple interpretations of claims. Another avenue of research is required to incorporate the claimant's intent: Satirical content, while most often being ❌ Compromised, could be considered ✅ Intact, given its humorous intention. Finally, since VERITAS only evaluates verdict scores, future work would need to develop evaluation approaches for the presented evidence.

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. While the share of duplicates in the release data appears to be very low (App. G), semantically equivalent claims appear across languages, particularly when fact-checking organizations publish parallel versions of the same claim (e.g., AFP). While this reflects real-world dissemination, it introduces a minor redundancy within splits. Future VERITAS iterations should introduce textual and visual claim deduplication for sampling.

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. While analysis shows no noticeable advantage on rectified claims (App. F), explicit prompting or finetuning strategies might still exploit shortcuts.

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 or fact-checking organizations themselves—could impair future data collection. While the pipeline has operated reliably for data from the past six years and successfully demonstrated its extension in Q1 2026, its long-term robustness cannot be guaranteed.

Finally, although fully automated, the construction of VERITAS incurs non-trivial API and computational costs. Creating the full benchmark required approximately $16.5 K so far, corresponding to about $660 per quarter split. Additionally, hosting LLAMA 4 MAVERICK in-house required roughly 2.9 K GPU hours on 8 NVIDIA H100 GPUs, i.e., about 116 h per quarter. Note that these costs do not affect users of VERITAS, which is freely available at veritas.mai.informatik.tu-darmstadt.de.

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, when used to monitor social networks, 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, for copyright compliance, and to avoid automated scraping, access to VERITAS is restricted to eligible researchers via a gated request process. 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.

Acknowledgments

We thank **Tobias Wieczorek**, **Florian Schröter**, **Aritra Marik**, and the other evaluators for their great work, each annotating dozens of claims and, thus, enabling the validation of VERITAS.

We gratefully acknowledge support from the hessian.AI Service Center (funded by the German **Federal Ministry of Research, Technology and Space (BMFTR)**, grant no. **16IS22091**) and the **hessian.AI Innovation Lab** (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. **S-DIW04/0013/003**).

The research was partially funded by a LOEWE Start-Professur (**LOEWE/4b//519/05.01.002-(0006)/94**), LOEWE-Spitzen-Professur (**LOEWE/4a//519/05.00.002-(0010)/93**) and an Alexander von Humboldt Professorship in Multimodal Reliable AI, sponsored by the

BMFTR and has benefited from the Excellence Cluster "Reaonsable AI" by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) under Germany's Excellence Strategy – **EXC-3057**.

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. Nonprobative 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

Tab. 4 summarizes the LLMs used in the VERITAS construction pipeline. At the time of processing Q1 2020 – Q4 2025, rate limits were too restrictive for the newly introduced GEMINI 3 family, therefore we defaulted to GEMINI 2.5 PRO and GEMINI 2.5 FLASH. We upgraded to GEMINI 3.1 PRO and GEMINI 3 FLASH for Q1 2026.

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 3.1 PRO	5, 6, 7	`gemini-3.1-pro-preview`
GEMINI 2.5 FLASH	5, 7	`gemini-2.5-flash`
GEMINI 3 FLASH	5, 7	`gemini-3-flash-preview`
CLAUDE SONNET 4.5	6	`claude-sonnet-4-5-20250929`
LLAMA 4 MAVERICK	6	`Llama-4-Maverick-17B-128E-Instruct-FP8`

Table 4: The Large Language Models (LLMs) used in this work, along with the version specifier and the corresponding VERITAS pipeline stage where they were used. Stages with GEMINI 2.5 models were upgraded to the GEMINI 3 family after it became available.

B VERITAS Stage Implementation Details

Stage 1: Review Discovery ClaimReviews are obtained via the Google Fact Check Tools API[3] and DataCommons[4]. Extracted review instances are refreshed by re-downloading ClaimReview data directly from the publisher websites. Language is determined using `langdetect`[5].

We observed several reviews that indicate a claim date after the review's publication date. Since it is not possible to fact-check "future" claims, we assume this is an error in the metadata provided by the publishing fact-checking organization. Therefore, we dismiss all of these 1.2 K (0.3%) reviews.

Stage 2: Publisher Identification We consider a fact-checking organization as credible if it is a signatory of the International Fact-Checking Network (IFCN)[6] or member of the European Fact-Checking Standards Network (EFCSN)[7].

Stage 3: Article Scraping We scrape article content for each review using a dynamic engine based on FIRECRAWL[8] and DECODO[9]. Missing review publication and modification dates are obtained via HTML metadata.

Stage 4: Appearance Retrieval Appearances are scraped via `scrapeMM`[10] which downloads media and text through social media APIs and via HTTP requests using `yt-dlp`[11], FIRECRAWL, and DECODO.

Stage 6: Verdict Standardization To enable the ensemble LLMs to meaningfully return scoring values on the $[-1, 1]$ interval, we discretize it into 7 bins as detailed in Fig. 6.

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 Score Discretization

Instead of a finite set of labels, VERITAS uses continuous scores to rate claims and media w.r.t. their specific properties. Scores range from -1 (the ✗ Negative extreme) to 1 (the ✓ Positive extreme), where values in between indicate different degrees of uncertainty. This allows the use of distance-based metrics and straightforward ensemble rating aggregation.

However, when computing accuracy or using LLMs to predict ratings, the scale must be discretized. We use two different discretization strategies: 3-bin discretization (representing the conventional classification approach) and a more fine-grained 7-bin approach that accounts for uncertainty. We use the 3-bin solely for reporting accuracy, while the 7-bin is applied to LLM predictions. Please refer to Fig. 6 for the exact scale-to-label binning mapping.

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

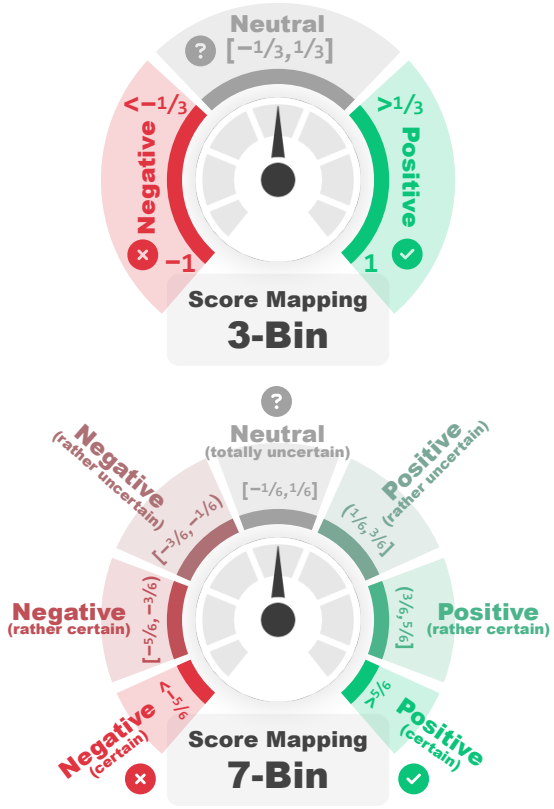


Figure 6: Discretization strategies for mapping from the continuous rating interval $[-1, 1]$ to 3 or 7 discrete labels, respectively.

D On the Choice of the Metrics

We report four metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy (3-bin and 7-bin). Since VERITAS operates on a continuous label space, error-based metrics are more suitable to capture graded deviations from the ground truth.

We adopt MSE as the primary metric for performance evaluation for two reasons. First, unlike Accuracy, MSE accounts for the magnitude of errors. Predictions close to the ground truth incur only a small penalty, whereas Accuracy assigns the same penalty to all misclassifications regardless of distance (cf. Fig. 7). Second, compared to MAE, MSE penalizes large deviations more strongly: Confusing opposite extremes (e.g., ✓ True vs. ✗ False) incurs a substantially higher penalty than small errors. This behavior better reflects the practical cost of severe misjudgments while remaining more forgiving for near-correct predictions. Table 5 shows how to interpret the MSE values in the context of VERITAS.

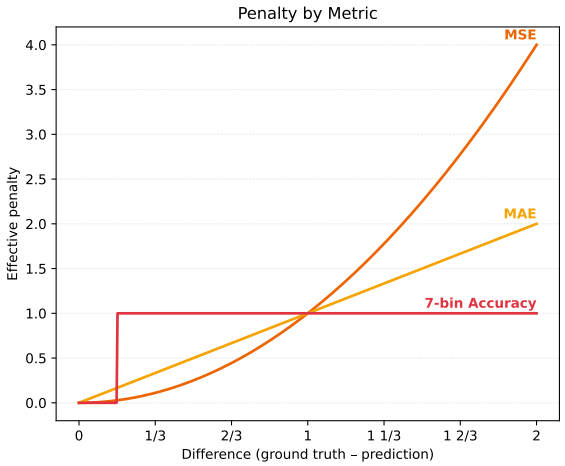


Figure 7: Comparison of metrics used for VERITAS. Lines indicate the effective penalty on the final evaluation score. **MSE** = Mean Squared Error (main metric of VERITAS), **MAE** = Mean Absolute Error.

E VERITAS Release 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). Additionally, original ✓ Intact claims take precedence over rectified claims.

F Analysis of Potential Stylistic Shortcuts in Rectified Claims

A potential concern is that rectified claims may contain stylistic artifacts that models could exploit as shortcuts, rather than relying on factual reasoning. To assess this empirically, we conducted an additional experiment evaluating GPT-5.2 and GEMINI 3 PRO on three subsets: (i) all 147 ✓ Intact, original claims, (ii) 300 ✓ Intact, rectified claims, and (iii) 300 ✗ Compromised, original claims.

The results (depicted in Tab. 6) provide no strong evidence for a systematic shortcut effect. While

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 5: 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).

	✓ Intact (original)	✓ Intact (rectified)	✗ Compr. (original)
GPT-5.2			
N Claims	147	300	300
MSE ↓	0.577	0.460	0.189
MAE ↓	0.450	0.423	0.233
3-bin Acc. ↑	71.4	66.3	82.0
GEMINI 3 PRO			
N Claims	147	300	300
MSE ↓	0.264	0.270	0.349
MAE ↓	0.264	0.264	0.336
3-bin Acc. ↑	88.4	83.7	74.7

Table 6: Performance comparison on original vs. rectified claims for intact and compromised subsets, Accuracy in %.

GPT-5.2 exhibits a moderate improvement on rectified intact claims over original intact ones (MSE difference of approximately 0.12), performance remains far from trivial, with substantial residual errors. In contrast, GEMINI 3 PRO shows no consistent advantage for rectified claims; performance is comparable or slightly worse than on original claims.

Overall, these findings suggest that, even if minor stylistic signals are present, they do not translate into a reliable or substantial shortcut across models. However, this analysis does not preclude the possibility that such patterns could be exploited under targeted prompting or finetuning, which we leave for future investigation.

G Analysis of Duplicates in VERITAS

A potential concern is the presence of duplicate or highly similar claims, which could reduce effective diversity. To quantify redundancy in VERITAS, we computed pairwise cosine similarity over all 25 K claim text embeddings using OpenAI's `text-embedding-3-large` model.

The results in Tab. 7 indicate that redundancy is limited. At most 0.45% of claims exhibit substantial overlap in their textual assertions (cosine similarity ≥ 0.95), while near-identical claims (cosine similarity ≥ 0.99) account for only 0.1%. Importantly, these estimates represent a worst-case scenario based solely on textual similarity. When additionally considering differences in associated media or claim publication dates, the effective redundancy is further reduced.

Overall, this analysis suggests that VERITAS maintains a high degree of diversity, with only minimal duplication.

H Sensitivity to Proprietary LLM Backbones

Dependence on proprietary models raises questions regarding reproducibility. To empirically assess the sensitivity of VERITAS to the choice of backbone, we partially regenerated Q4 2025 using a substantially different LLM family.

Specifically, we replaced GPT-5.2 with GEMINI 3.1 PRO and GPT-5.2 MINI with GEMINI 3 FLASH for all non-video claims. For claims involving videos, we substituted GEMINI 2.5 PRO with GEMINI 3.1 PRO and GEMINI 2.5 FLASH with GEMINI 3 FLASH. These changes affect stages 5 (Claim Normalization), 6 (Verdict Standardization), and 7 (Claim Rectification), while the ensemble used for verdict prediction in stage 6 remained unchanged.

In total, we regenerated 689 claims and compared them against the original pipeline outputs. Manual inspection shows that regenerated claims preserve the same core assertions, with only minor differences in wording or level of detail. Quantitatively, comparing ensemble-generated verdicts on

Cos-sim	Typical Observation	# Claim Pairs	Worst-case Share
≥ 0.90	Same topic, but different assertions	280	1.12%
≥ 0.95	Mostly same assertions, minor wording differences	112	0.45%
≥ 0.99	Nearly identical (1–2 word or punctuation differences)	24	0.10%

Table 7: Redundancy analysis based on cosine similarity of claim text embeddings.

original versus regenerated claims yields an MSE of 0.0217 across all low-level properties and an MSE of 0.0069 for Integrity.

These results indicate extremely high agreement and suggest that the pipeline is largely invariant to the specific proprietary backbone employed. While this does not yet establish full reproducibility with open-weight models, it provides evidence that VERITAS is not tightly coupled to a single model family. As open-weight multimodal models continue to improve, future iterations of the pipeline can replace individual components accordingly.

I VERITAS Statistics

I.1 ClaimReview

We obtained a total of $398,327\,K$ 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 metadata 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. Fig. 8 shows the number of reviews returned by the three different sources. Fig. 9 displays the number of reviews per quarter since the introduction of ClaimReview in 2016. Fig. 10 shows the most common verdict labels as provided by the fact-checkers.

I.2 Appearances

We identified 495,642 appearances. Fig. 11 depicts the shares of appearances, containing the URL to the original source and/or to the archived record. In about 92% of the cases, we were able to identify the original appearance URL, only for 8% of the appearances we did not. 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

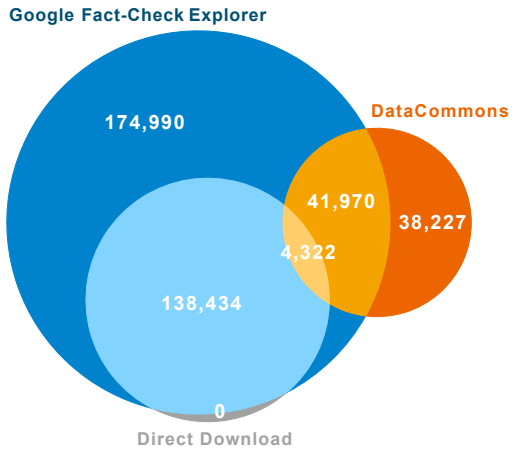


Figure 8: Origins of ClaimReviews as obtained by stage 1.

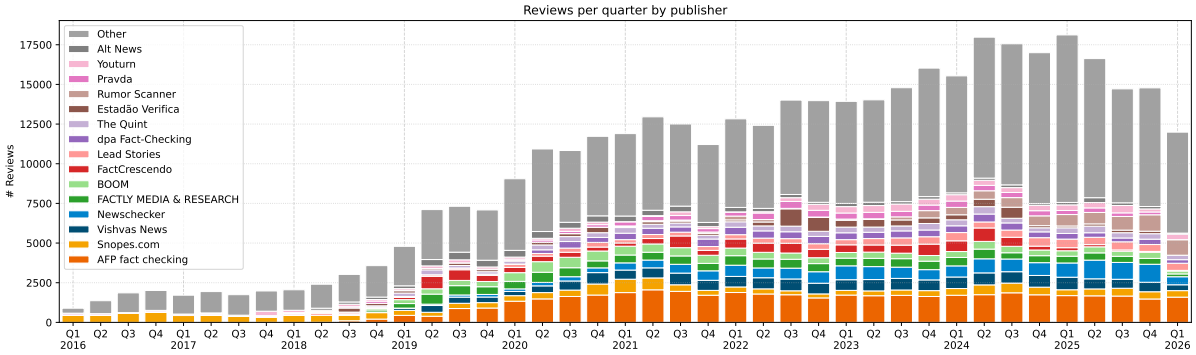
appearances occur are listed in Fig. 12. The distribution of archiving services is shown in Fig. 13. The most prominent are Perma.cc (41.8%) and Archive.today (38.2%). The quarterly shares of appearances by platform are shown in Fig. 18b.

I.3 Claims

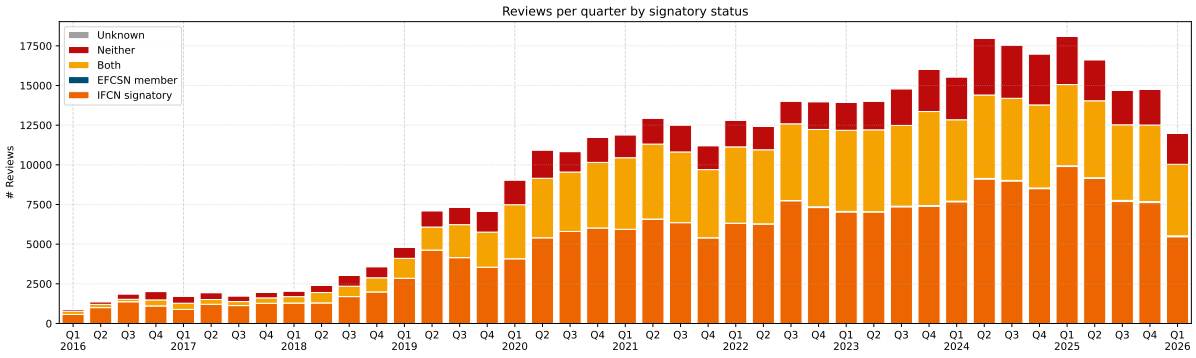
A total of $86,372\,K$ claims (original, rectified, and dismissed) were obtained by stage 5. The Figures 14 and 15 show the quarterly claim distributions for several different properties for the **released** $24\,K$ claims. The statistics for original (i.e., non-rectified and dismissed) claims are depicted in Figures 17 and 18.

I.4 Verdicts

Tab. 8 shows the total number of labels after discretizing scores into 3 bins as described in App. C. 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. Table 8 and Figure 19 show the distribution of scores for all properties.



(a) ClaimReviews by publishing organization.



(b) ClaimReviews by publisher signatory status.

Figure 9: ClaimReview statistics, as obtained by stages 1 and 2, showing *all* ClaimReviews exposed by Google and Data Commons retrieved by Dec 31, 2025.

Figure 10: Raw ratings as provided by ClaimReview for English reviews.

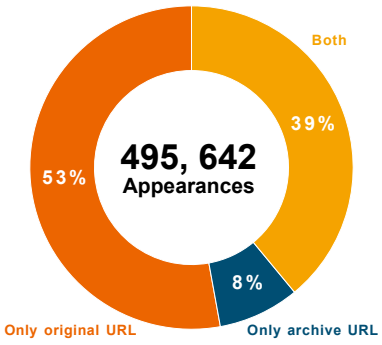


Figure 11: Distribution of appearances that have the original source URL, an archive URL, or both.

I.5 Dismissal Reasons

The VERITAS pipeline can discard reviews and claims for various reasons. Fig. 20 depicts the 10 most common reasons.

J Baseline Experiments

J.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 temporal filtering (only content published before the claim date is retained) to avoid temporal leakage. For the top 3 filtered results, we attempt to scrape the full page content using `scrapeMM`; 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 Tab. 9 and Fig. 21. Figure 22 shows the results per modality.

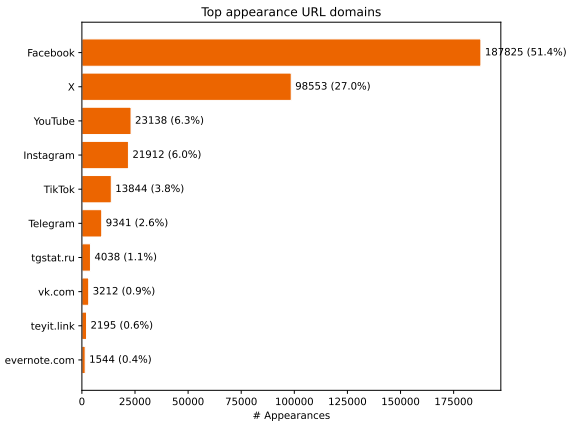


Figure 12: Appearances per platform, showing the top 10.

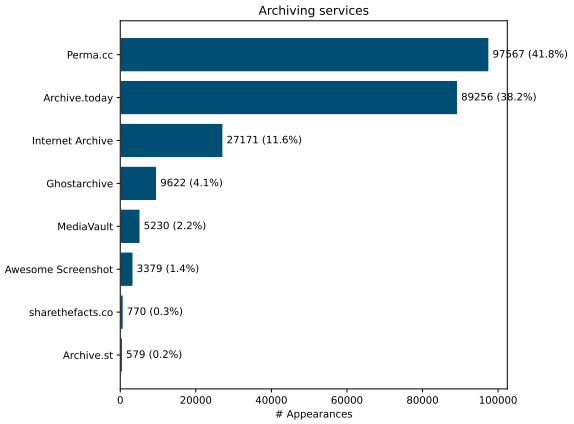


Figure 13: Appearances per archiving service.

K Human Evaluation

K.1 Setup Details

Annotation Organization. We conducted two phases of evaluation—both employing the practically same annotation process.

In the first phase, 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[12]. We dismissed annotations with score difference larger than 1, same for claims that received less than 2 annotations.

In the second phase, we extended the evaluation to a crowd of 46 people recruited via Prolific in order strongly to increase the number of evaluated claims. An updated tutorial video was presented to them [13], incorporating audience-adequate clarifications and explanation improvements learned from phase one. After watching the tutorial, evaluators were screened: They must correctly answer 10 out of 10 non-trivial multiple-choice questions in order to pass the screening. Additionally, we manually reviewed the quality of the annotations of each annotator. 24 annotators passed this screening. They yielded a total of 444 additional annotations.

Property	Negative	NEI	Positive
Authenticity	3,659	2,591	5,769
Contextualization	5,726	462	5,831
Veracity	6,539	129	12,610
Context Coverage	192	4	12,446
Integrity	12,453	94	12,453

Table 8: Decision counts for all five properties.

Method	Search	Error Rates (↓)		Accuracy (↑)	
		MSE	MAE	7-bin	3-bin
Gemini 3 Flash	-	0.334	0.248	**70.4**	89.7
Gemini 3.1 Pro	-	**0.309**	**0.235**	**70.3**	**90.3**
Claude Opus 4.6	-	0.363	0.375	41.0	76.0
GPT-4o	-	0.565	0.504	37.0	64.0
GPT-5.2	-	0.520	0.552	21.4	56.3
Llama 4 Maverick	-	0.735	0.622	24.6	58.1
Gemma 4 (31B)	-	0.525	0.374	59.0	82.2
Qwen 3.5 (397B)	-	0.600	0.440	50.9	74.4
Gemini 3 Flash	✓	**0.158**	**0.168**	**72.5**	**94.1**
Gemini 3.1 Pro	✓	0.234	0.222	68.0	88.5
Claude Opus 4.6	✓	0.305	0.358	42.3	72.6
GPT-4o	✓	0.609	0.502	42.5	65.7
GPT-5.2	✓	0.325	0.412	25.2	76.1
Llama 4 Maverick	✓	0.641	0.585	26.7	59.2
Gemma 4 (31B)	✓	0.265	0.243	65.0	89.2
Qwen 3.5 (397B)	✓	0.267	0.278	56.4	85.9

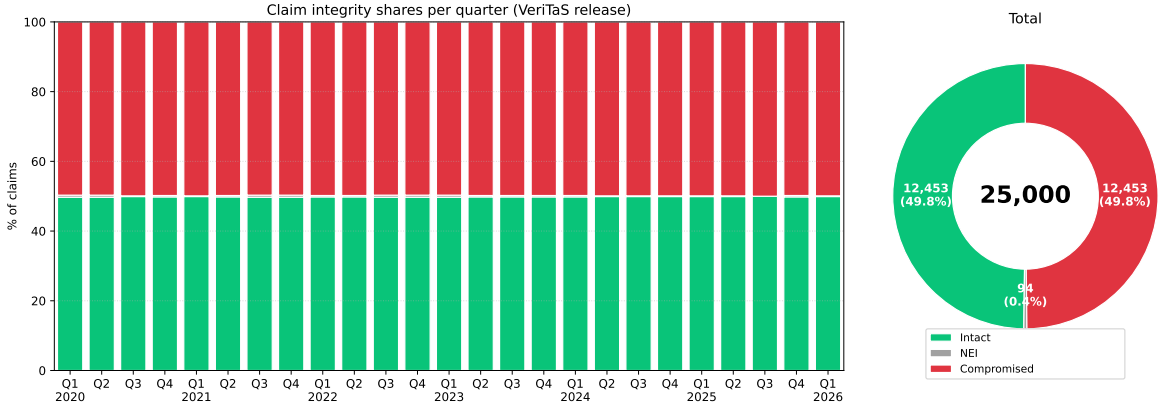
Table 9: Results on the **longitudinal data split**. Baselines have been tested once using parametric knowledge only (without search tool) and once using web search (with search tool), single runs.

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.

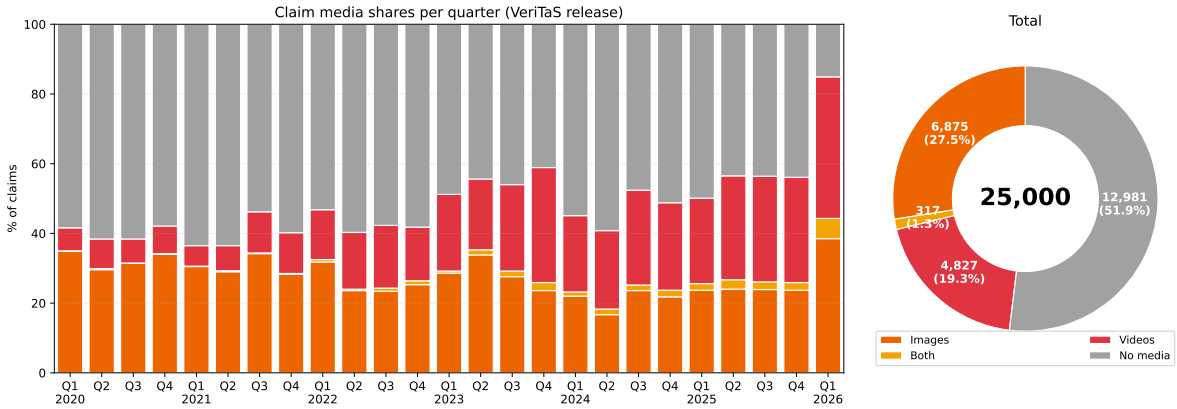
The evaluation then proceeds sequentially through the properties, following the same proce-

[12] Phase 1 annotation tutorial: youtu.be/ep9ssCf1ets

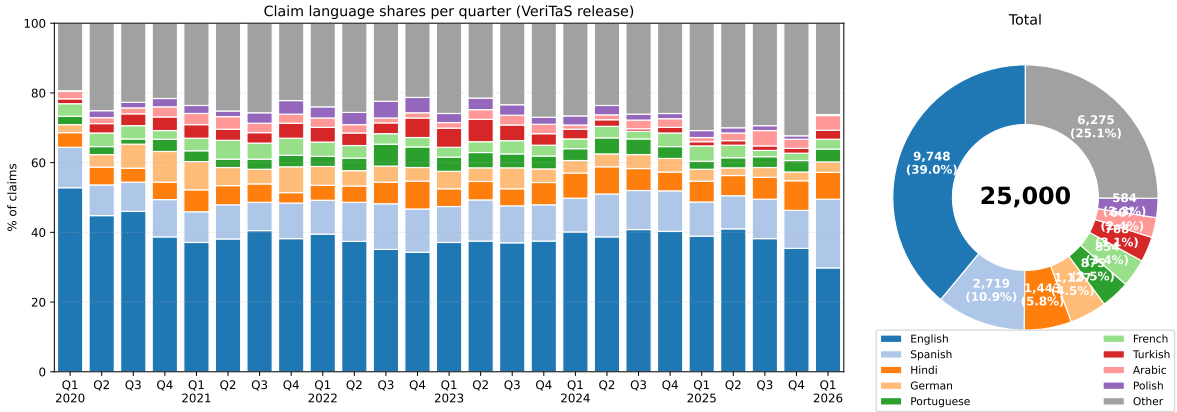
[13] Phase 2 annotation tutorial: youtu.be/lYAggVNrynY



(a) Integrity shares.

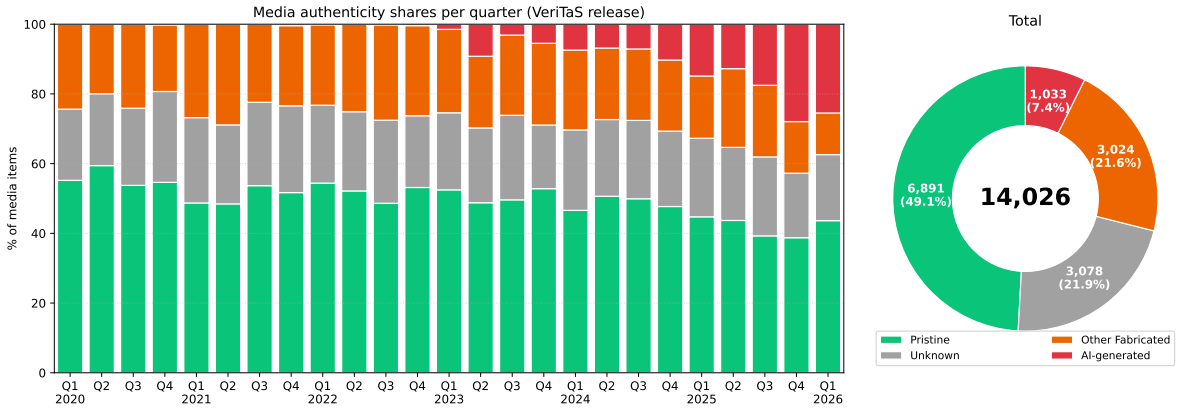


(b) Shares of claims featuring media. Note that in Q1 2026 the `scrapeMM` web scraping package was improved resulting in significantly higher media download success rates.

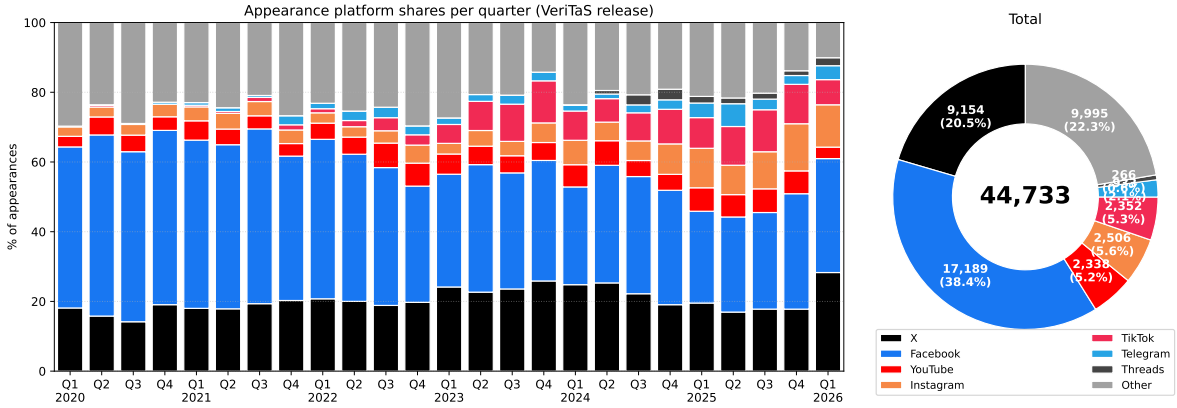


(c) Language shares.

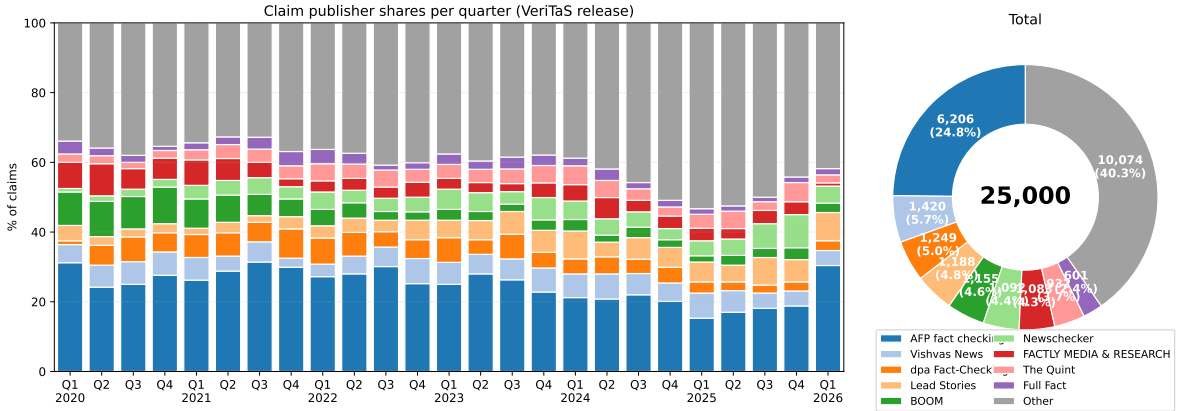
Figure 14: Statistics in the **released** VERITAS benchmark for all quarter splits.



(a) Media authenticity shares.



(b) Platform shares for all claim appearances.



(c) Publisher shares.

Figure 15: Statistics in the **released** VERITAS benchmark for all quarter splits.

(a) English (b) Spanish

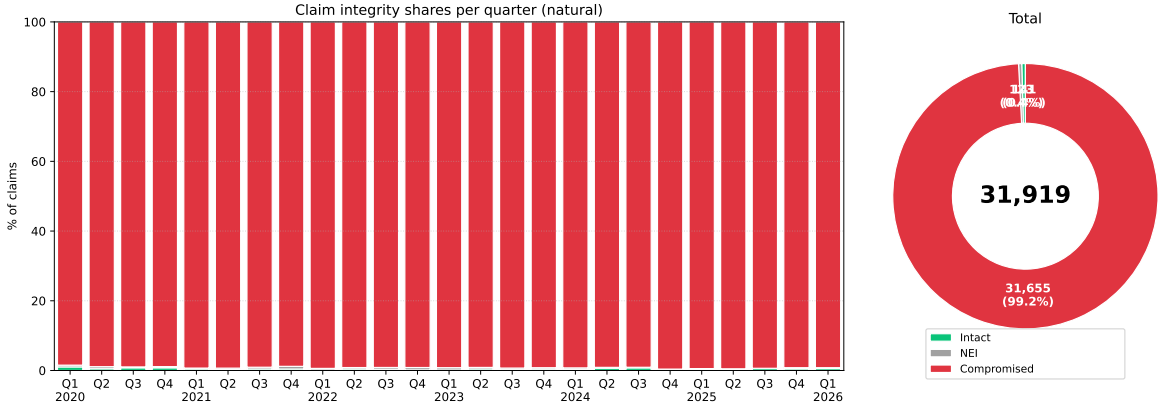
(c) Hindi (d) German

(e) Portuguese (f) French

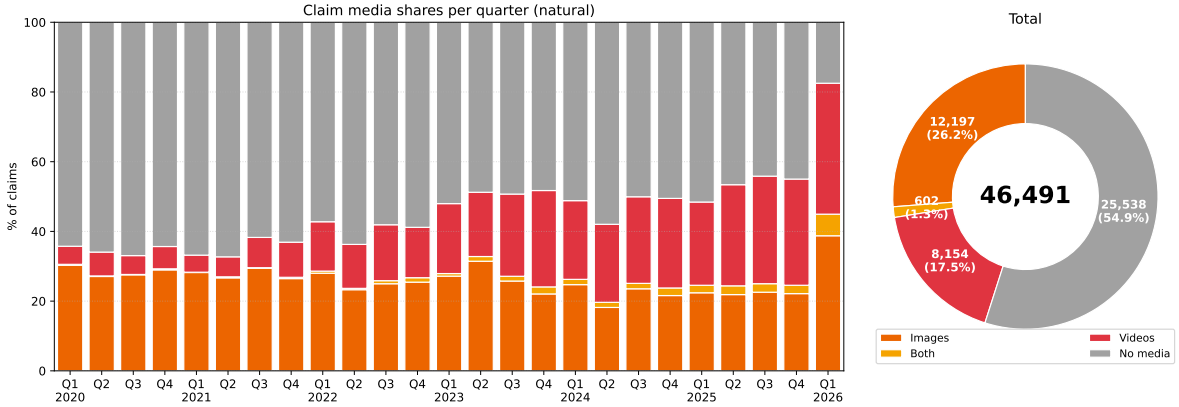
(g) Turkish (h) Russian

(i) Polish (j) Arabic

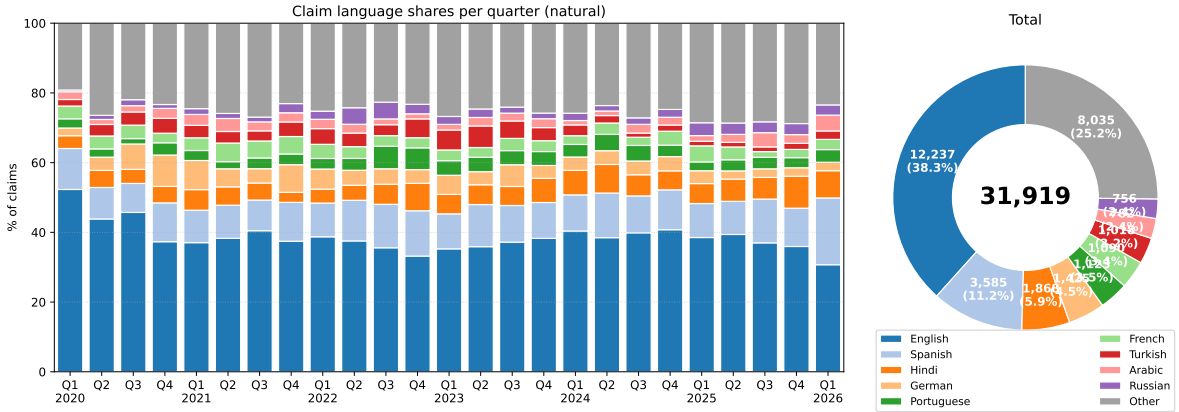
Figure 16: Common words occurring in VERITAS' claims for the ten most frequent languages.



(a) Integrity shares.

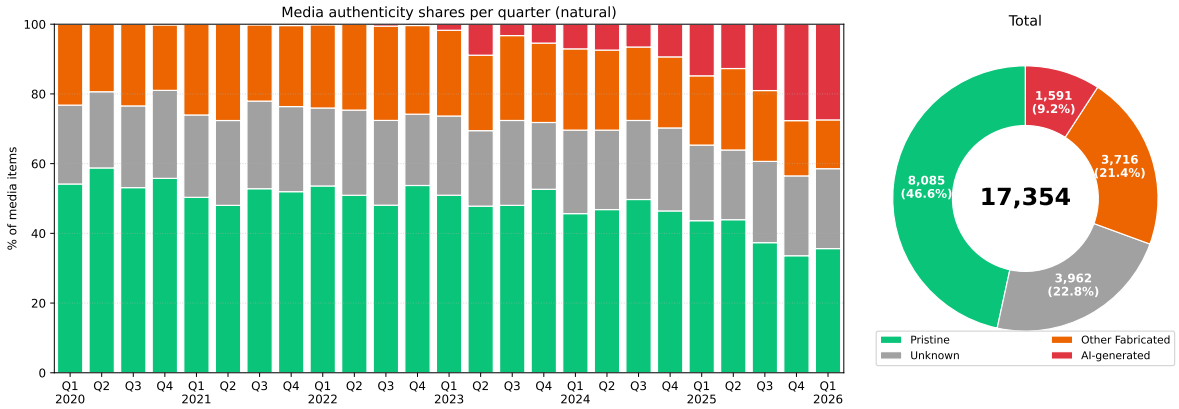


(b) Shares of claims featuring media. Note that in Q1 2026 the `scrapeMM` web scraping package was improved resulting in significantly higher media download success rates.

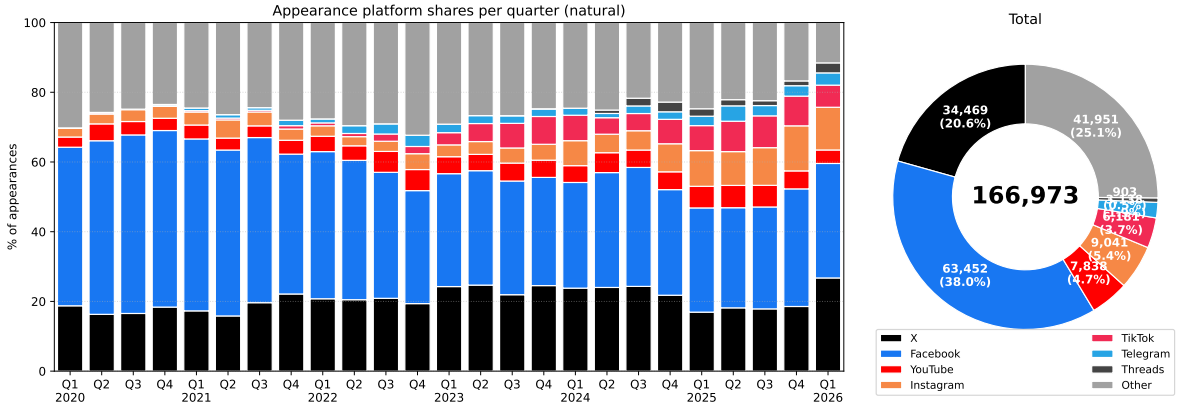


(c) Language shares.

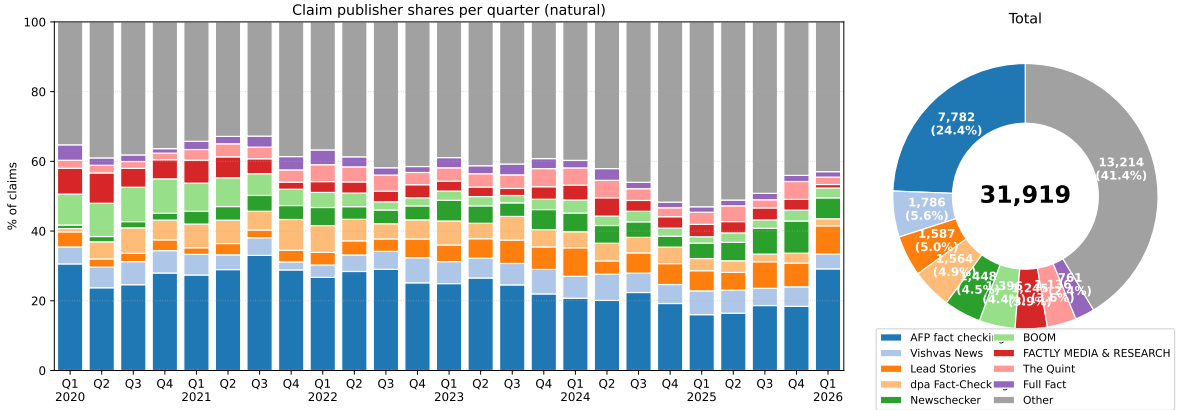
Figure 17: **Natural** data statistics as obtained before stage 7, i.e., before balancing/rectification and sampling.



(a) Media authenticity shares.



(b) Platform shares for all claim appearances.



(c) Publisher shares.

Figure 18: **Natural** data statistics as obtained before stage 7, i.e., before balancing and sampling.

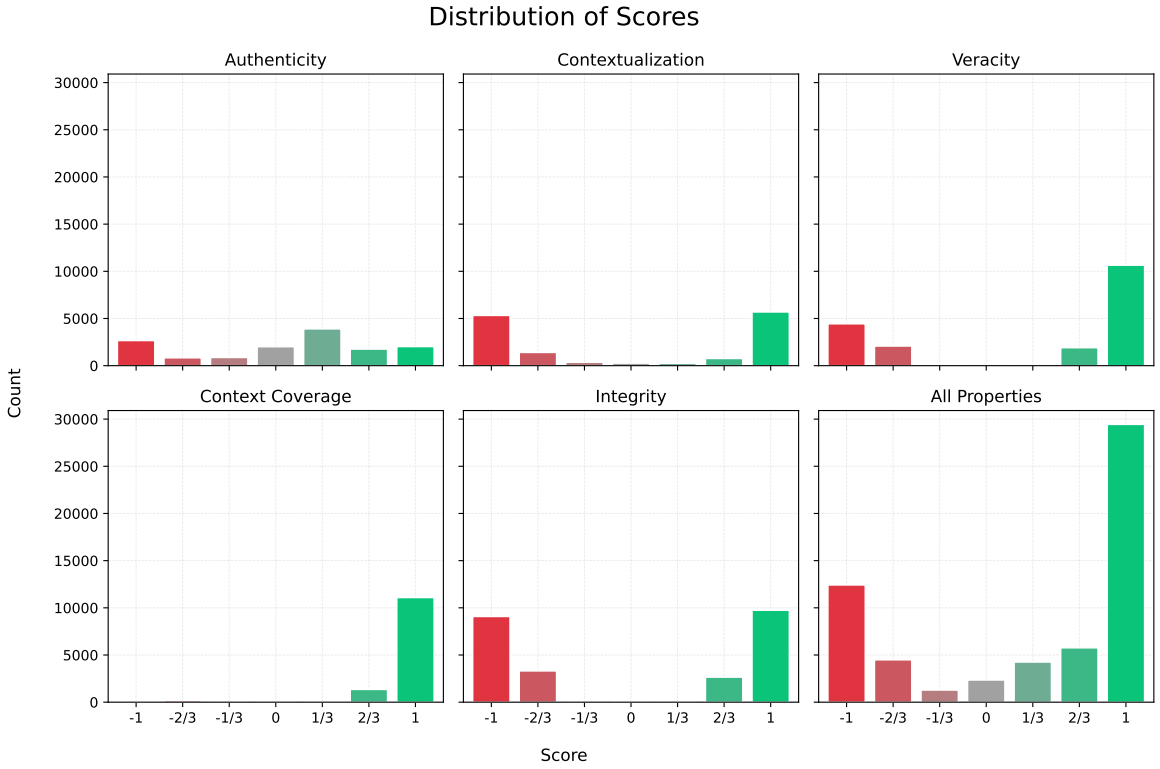


Figure 19: Score distribution of the released VERITAS benchmark for all five properties and their sum.

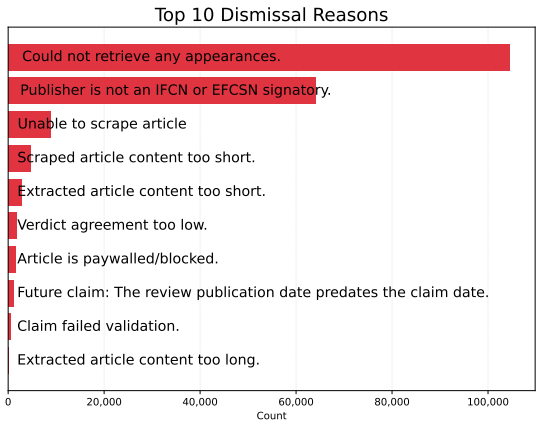


Figure 20: Most common reasons for a claim/review to be dismissed by the VERITAS pipeline

Property	N	**Error Rates (↓)**		**Accuracy (↑)**	
		MSE	MAE	7-bin	3-bin
Authenticity	173	0.285	0.371	37.0	72.8
Contextualization	173	0.121	0.152	74.0	91.3
Veracity	138	0.021	0.072	74.6	97.8
Context Coverage	116	0.030	0.085	77.6	99.1
Integrity	204	0.034	0.102	69.1	97.5
All Properties	600	0.128	0.184	64.2	89.0

Table 10: Human evaluation results over N claims. The scores reflect agreement between human annotation and automated VERITAS annotations. **MSE** = Mean Squared Error, **MAE** = Mean Absolute Error, **Accuracy** = share of matches of n-bin discretized scores (in %).

dure as stage 6 of the VERITAS pipeline, cf. Fig. 3. Refer to the screenshots in Figure 24 for a full annotation walkthrough.

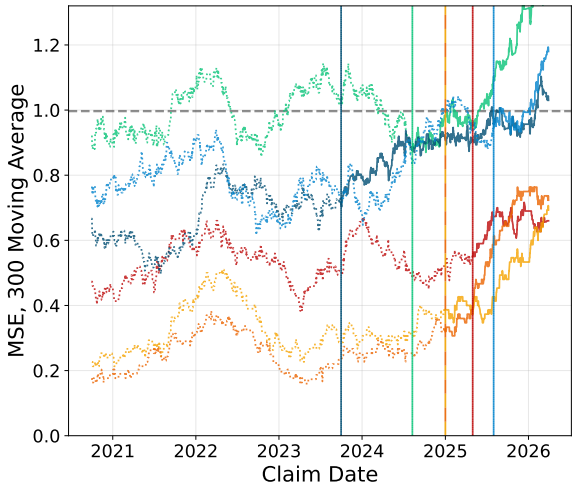
For each property, annotators provide a judgment on a seven-point scale incorporating the uncertainty values as defined in the 7 bin mapping, see Fig. 6. Each judgment requires a written explanation, informing our manual screening. When annotators select a confidence level below "rather certain," they must additionally describe the reason for their uncertainty.

K.2 Additional Results

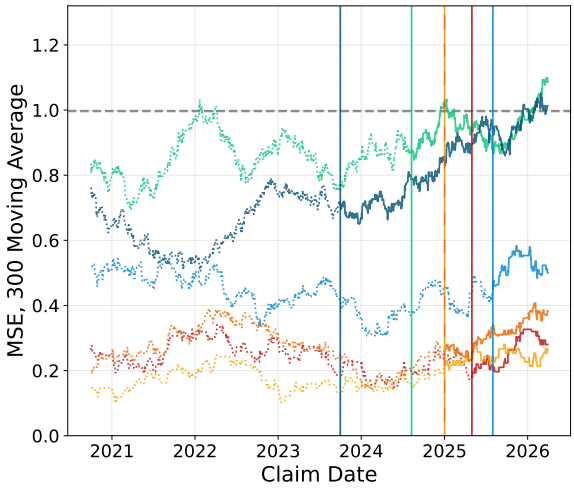
Tab. 10 and Fig. 25 compare human annotation with automated annotation for all 5 properties, including an aggregated statistic averaging all properties.

K.3 Discussion of Human Evaluation Disagreements

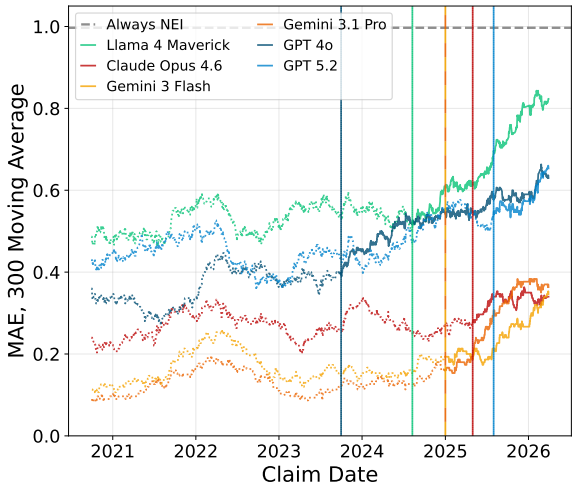
While error rates for *Media Authenticity* and *Media Contextualization* are relatively high, these do not directly translate to errors in the overall *Integrity*



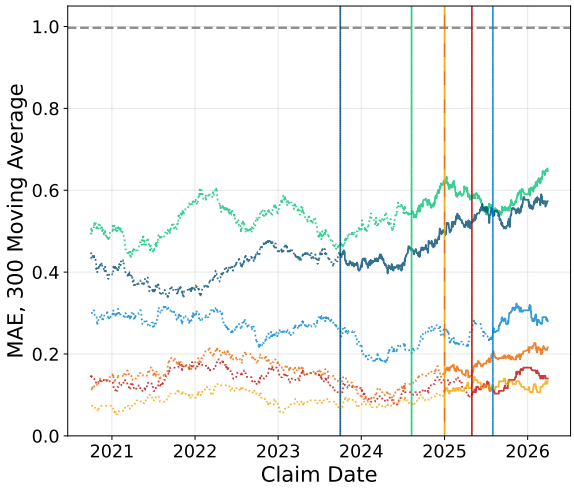
(a) MSE **without** web search.



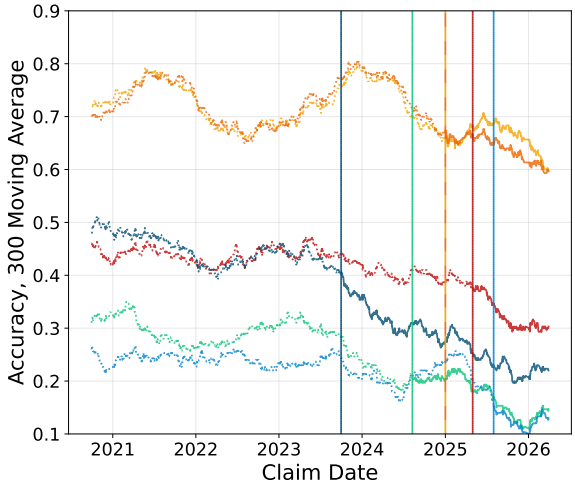
(b) MSE **with** web search.



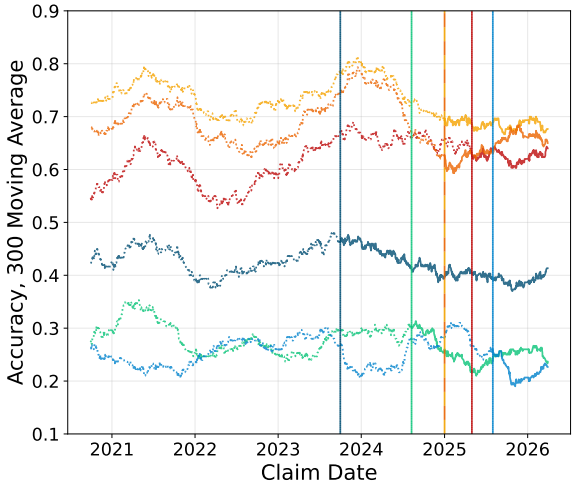
(c) MAE **without** web search.



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

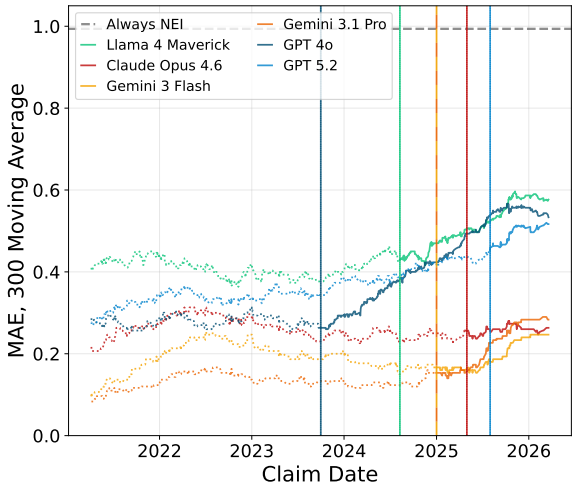


(e) Accuracy **without** web search.

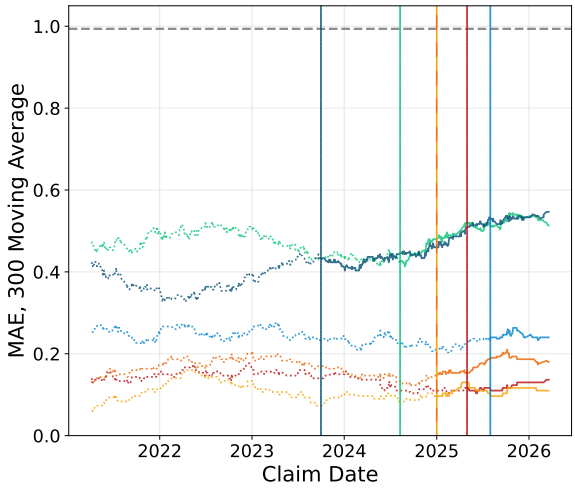


(f) Accuracy **with** web search.

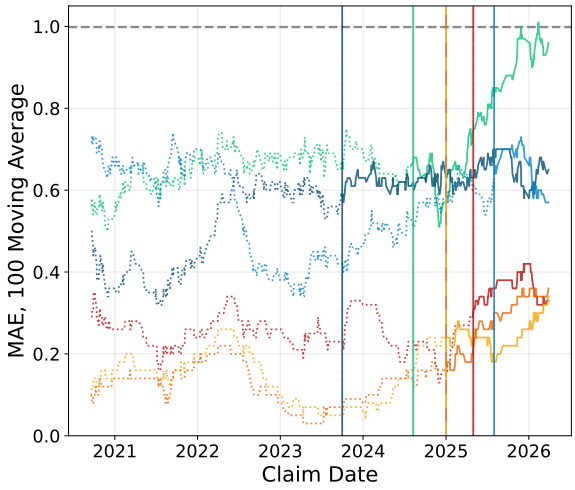
Figure 21: Baseline results on the **longitudinal split** for all three metrics Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy (by 3-bin discretization). All plots use a 200-claim moving average window. Vertical lines indicate knowledge cutoff dates.



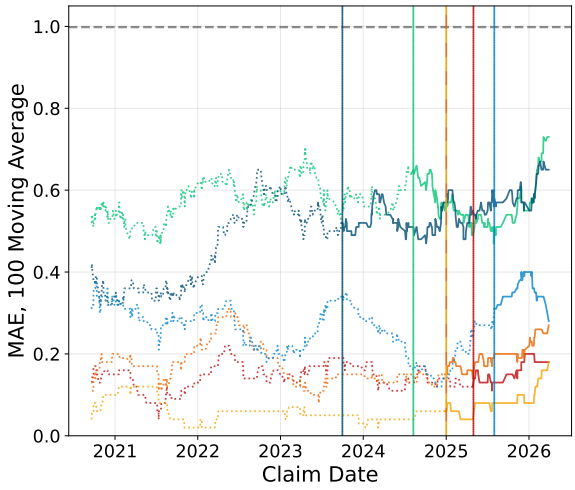
(a) MAE for text-only claims **without** web search.



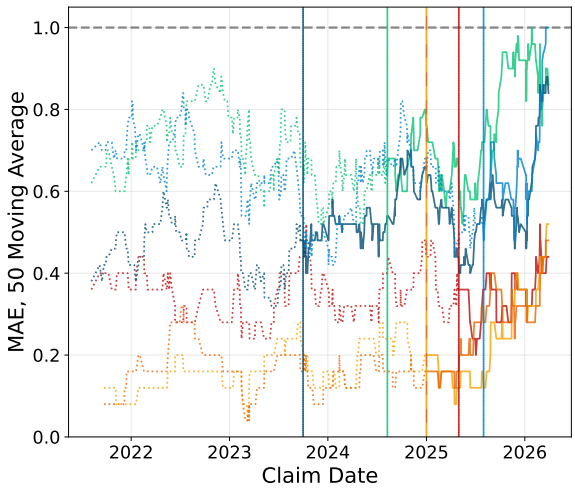
(b) MAE for text-only claims **with** web search.



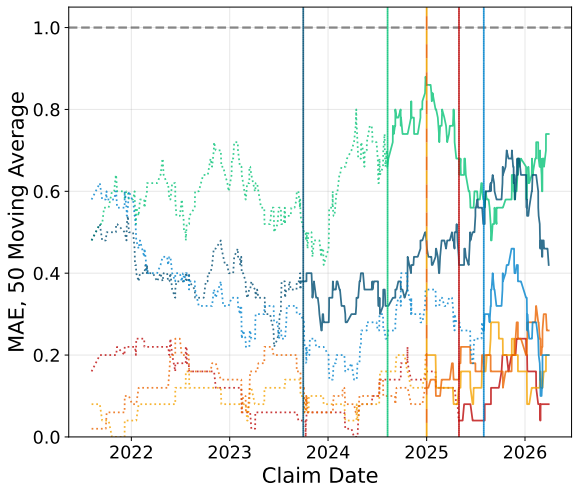
(c) MAE for text+image claims **without** web search.



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

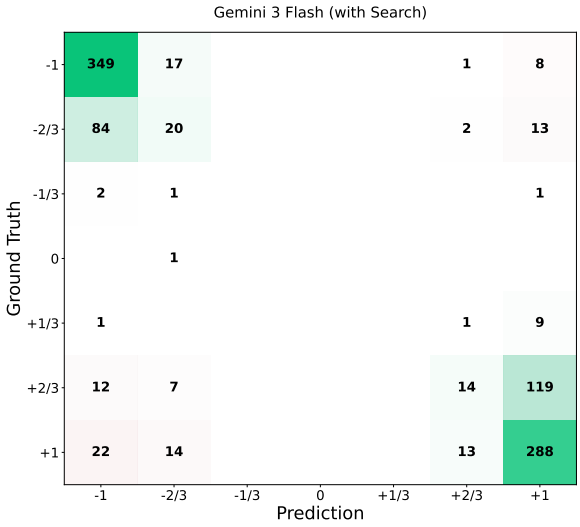


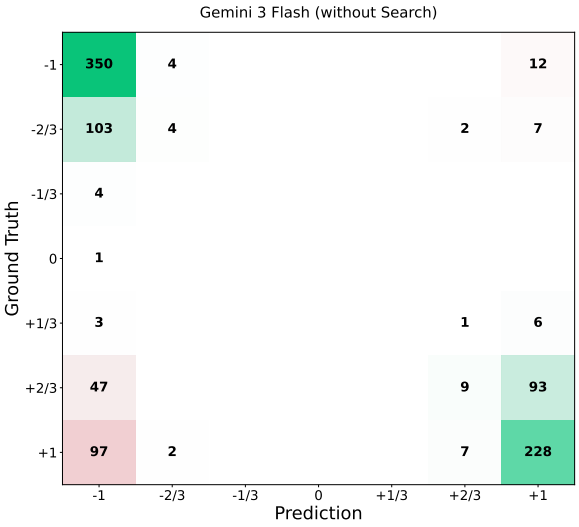
(e) MAE for text+video claims **without** web search.

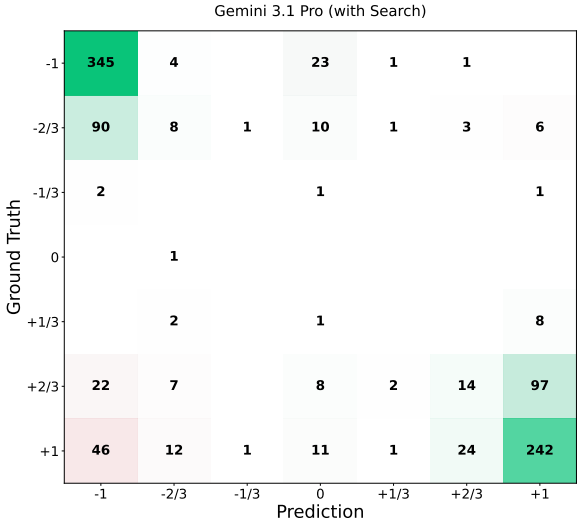


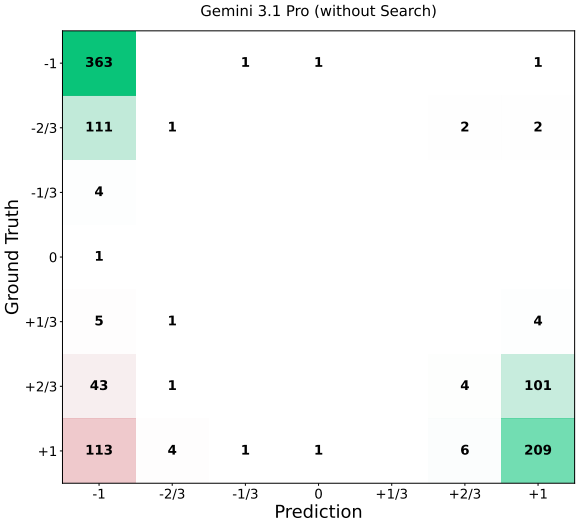
(f) MAE for text+video claims **with** web search.

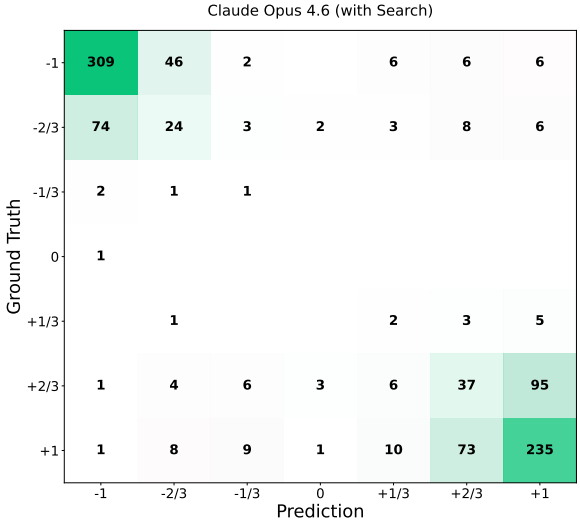
Figure 22: Baseline results on the **longitudinal split** for modality specific Mean Absolute Error (MAE). All plots use a moving average window. Vertical lines indicate knowledge cutoff dates.


Gemini 3 Flash (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 349 17 1 8
-2/3: 84 20 2 13
-1/3: 2 1 1
0: 1
+1/3: 1 1 9
+2/3: 12 7 14 119
+1: 22 14 13 288


Gemini 3 Flash (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 350 4 12
-2/3: 103 4 2 7
-1/3: 4
0: 1
+1/3: 3 1 6
+2/3: 47 9 93
+1: 97 2 7 228


Gemini 3.1 Pro (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 345 4 23 1 1
-2/3: 90 8 1 10 1 3 6
-1/3: 2 1 1
0: 1
+1/3: 2 1 8
+2/3: 22 7 8 2 14 97
+1: 46 12 1 11 1 24 242

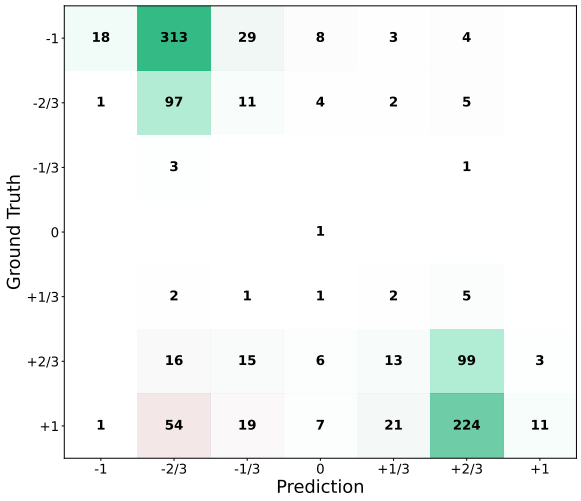

Gemini 3.1 Pro (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 363 1 1 1
-2/3: 111 1 2 2
-1/3: 4
0: 1
+1/3: 5 1 4
+2/3: 43 1 4 101
+1: 113 4 1 1 6 209

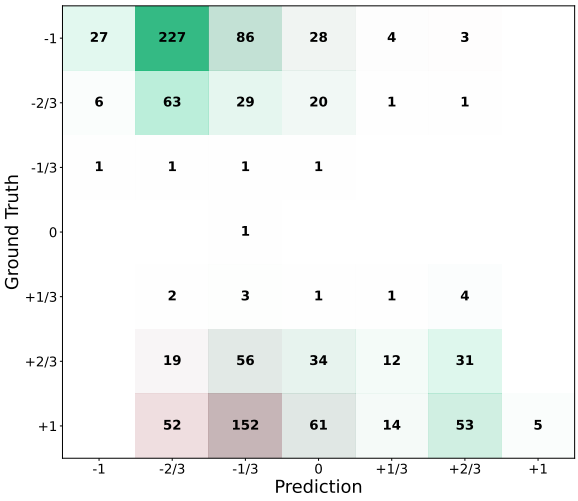

Claude Opus 4.6 (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 309 46 2 6 6 6
-2/3: 74 24 3 2 3 8 6
-1/3: 2 1 1
0: 1
+1/3: 1 2 3 5
+2/3: 1 4 6 3 6 37 95
+1: 1 8 9 1 10 73 235

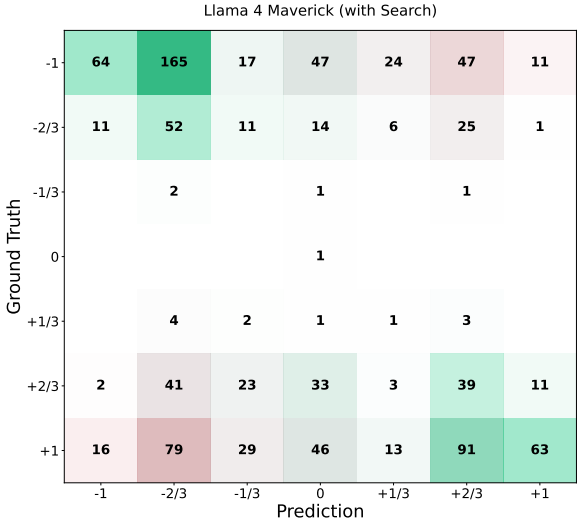

Claude Opus 4.6 (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
-1: 132 156 41 5 22 15 4
-2/3: 25 54 19 3 12 6 1
-1/3: 1 1 2
0: 1
+1/3: 3 1 2 3 2
+2/3: 27 15 3 29 56 22
+1: 3 34 34 5 70 135 56

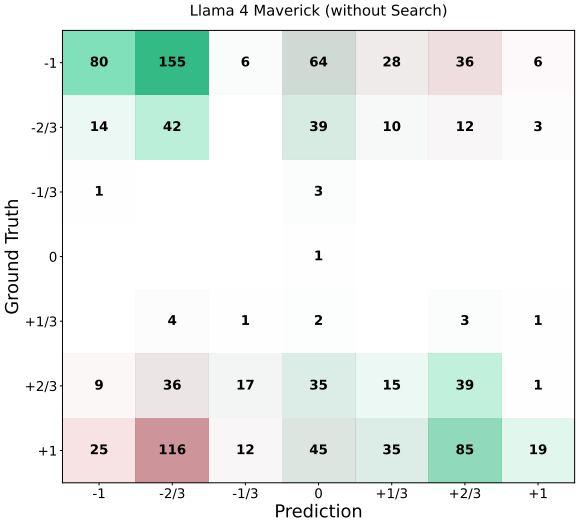

GPT 4o (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
219 83 18 32 1 11 11
43 39 8 18 7 5
1 2 1
1 2 2 3 2 2
47 21 6 27 7 20 24
82 33 1 55 1 59 106


GPT 4o (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
179 103 38 46 4 5
36 35 13 29 4 2 1
1 1 2
1 3 1 1 3 1 2
19 30 35 33 15 19 1
50 55 59 72 30 53 18


GPT 5.2 (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
18 313 29 8 3 4
1 97 11 4 2 5
3 1
1
2 1 1 2 5
16 15 6 13 99 3
1 54 19 7 21 224 11


GPT 5.2 (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
27 227 86 28 4 3
6 63 29 20 1 1
1 1 1 1
1
2 3 1 1 4
19 56 34 12 31
52 152 61 14 53 5


Llama 4 Maverick (with Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
64 165 17 47 24 47 11
11 52 11 14 6 25 1
2 1 1
1
4 2 1 1 3
2 41 23 33 3 39 11
16 79 29 46 13 91 63


Llama 4 Maverick (without Search)
Ground Truth
Prediction
-1 -2/3 -1/3 0 +1/3 +2/3 +1
80 155 6 64 28 36 6
14 42 39 10 12 3
1 3
1
4 1 2 3 1
9 36 17 35 15 39 1
25 116 12 45 35 85 19

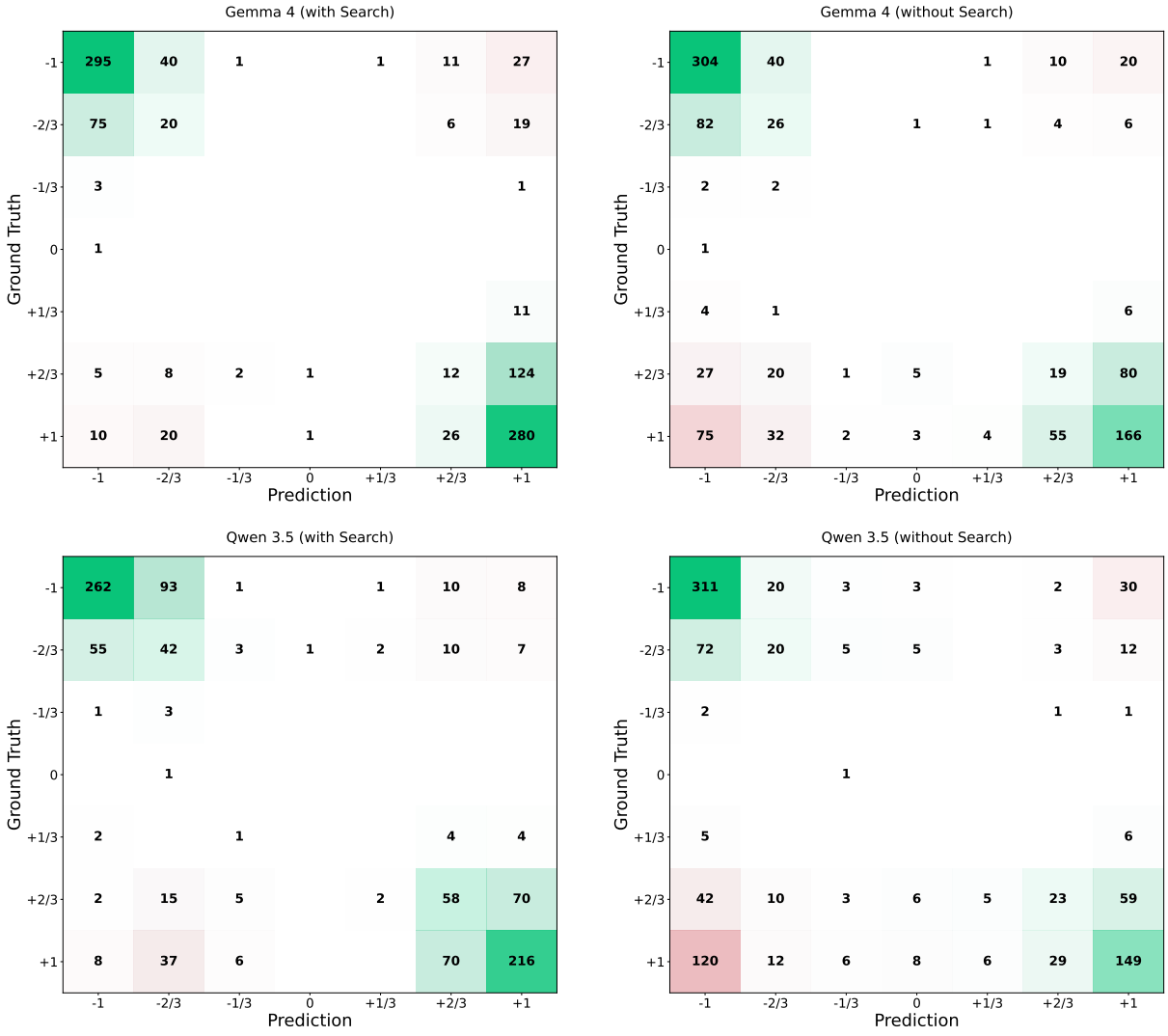
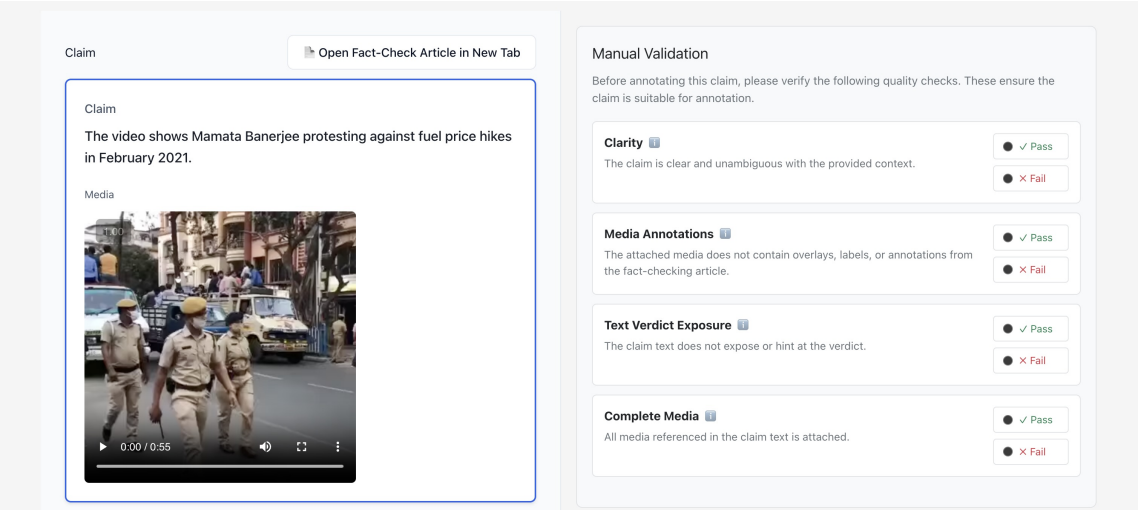
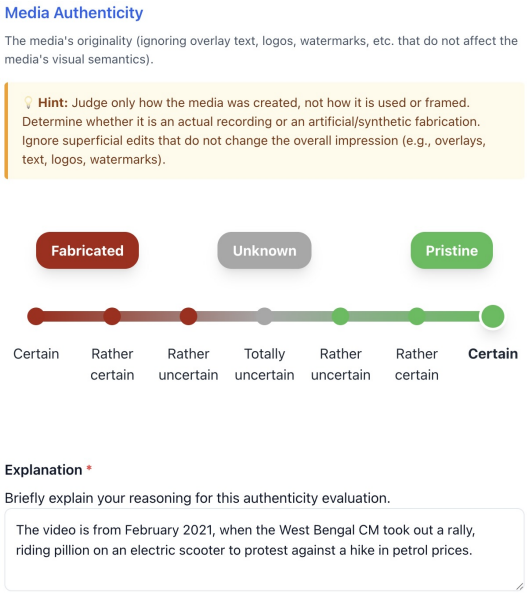


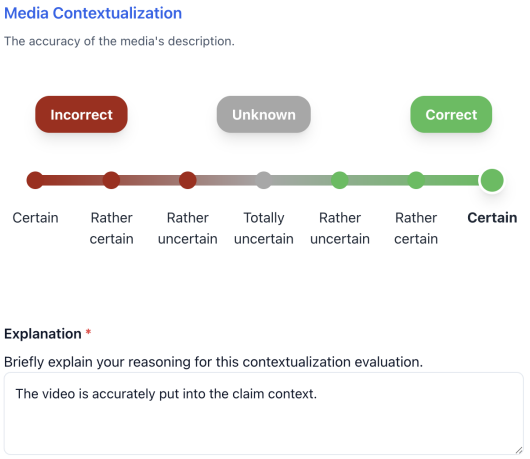
Figure 23: Confusion matrices for predicting Integrity for Q1 2026.



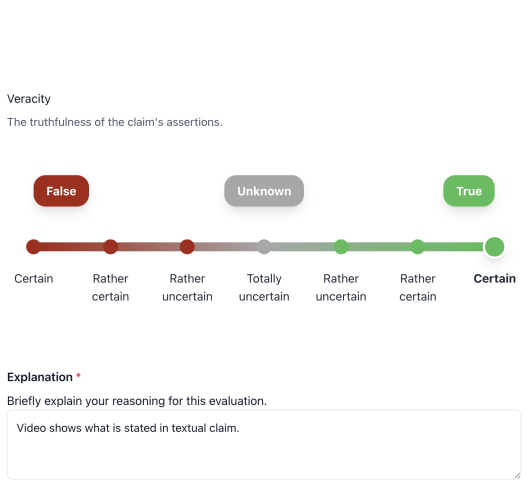
(a) Step 1: Validate that claim fulfills all quality requirements.



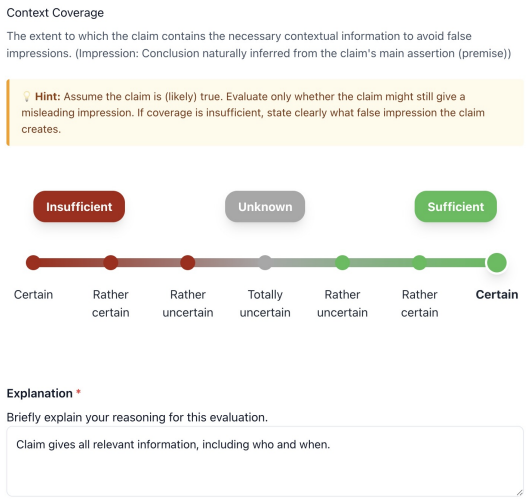
(b) Step 2: Assess media authenticity.



(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 24: 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).

assessment. Since *Integrity* aggregates multiple properties, disagreements at the property level do not necessarily imply incorrect end-to-end judgments. To better understand these discrepancies, we manually analyzed cases where human annotations and VERITAS predictions diverge.

Ambiguities in *Authenticity* Judgments. *Media Authenticity* exhibits numerous edge cases that make consistent annotation challenging for both humans and models. The boundary between Pristine and Fabricated media is often not clear-cut. For instance, benign modifications such as logos or watermarks are not considered manipulations, as they do not introduce misleading impressions. More challenging are staged scenes that are recorded without alteration: While the media itself is Pristine, correct contextualization requires revealing the artificial setup. Similarly, screenshots of user interfaces containing Fabricated images raise subtle questions. Under our definition, the digital environment is part of the real world, and thus such screenshots can be considered Pristine; however, this may be misleading when the claim's text refers to the fabricated content within the image rather than the screenshot itself. Related ambiguities arise for digital artwork, photographs of physical artwork, movies, or images containing textual overlays that convey false claims. These examples highlight that authenticity is inherently context-dependent and subject to interpretation. We leave further disambiguation to future work.

Cross-Property Interactions in *Contextualization*. We further observe that errors in *Media Contextualization* are often linked to interactions with textual claims. In cases where the media is only loosely related to the textual assertion (e.g., symbolic or illustrative images), the correct *Contextualization* label is NEI. However, VERITAS occasionally evaluates the textual claim during the *Contextualization* step, effectively refuting it instead of focusing on the media. While this behavior increases the measured error rate for *Contextualization*, it often still leads to a correct overall *Integrity* assessment. This suggests that some apparent errors arise from deviations in task decomposition rather than failures in factual reasoning. We will address this issue in future iterations of VERITAS.

L Claim Browser UI

To facilitate exploration of the VeriTaS dataset, we provide a Claim Browser that allows admitted researchers to browse, filter and inspect claims through a graphical user interface, see Fig. 26. It offers access to all gathered 104 K, not only the released, including also dismissed claims. Claims can be filtered along more than twenty dimensions, including free-text search over claim text, language, publication date range, verdict properties and validation status.

Selecting a claim opens a detailed view information on (1) the claim text, date, and associated media, (2) all model-individual label assignments with justifications and per-medium authenticity and contextualization scores, (3) appearances where the claim was observed, (4) link to the original fact-checker article, and (5) outcomes of the automated validation checks.

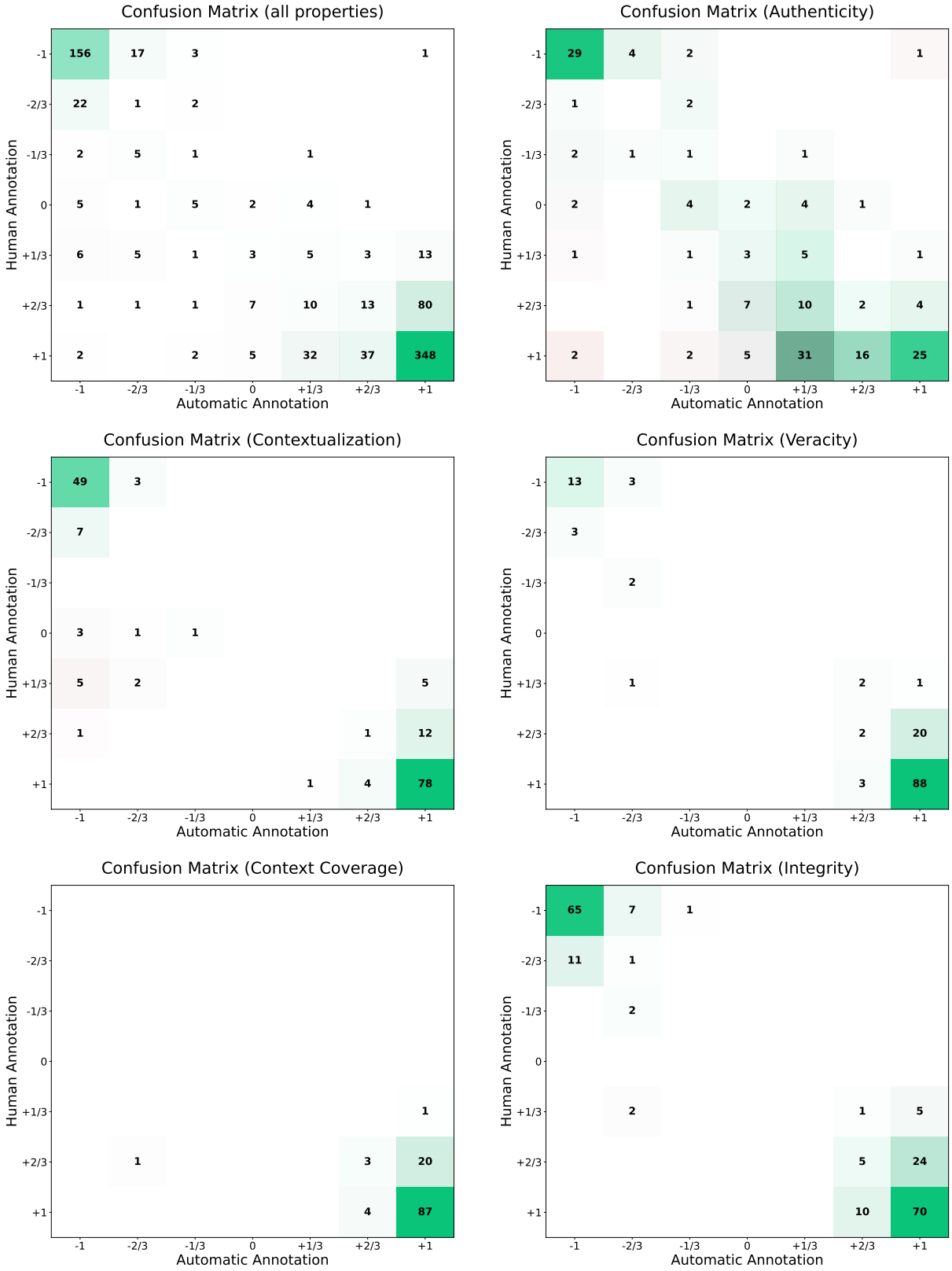
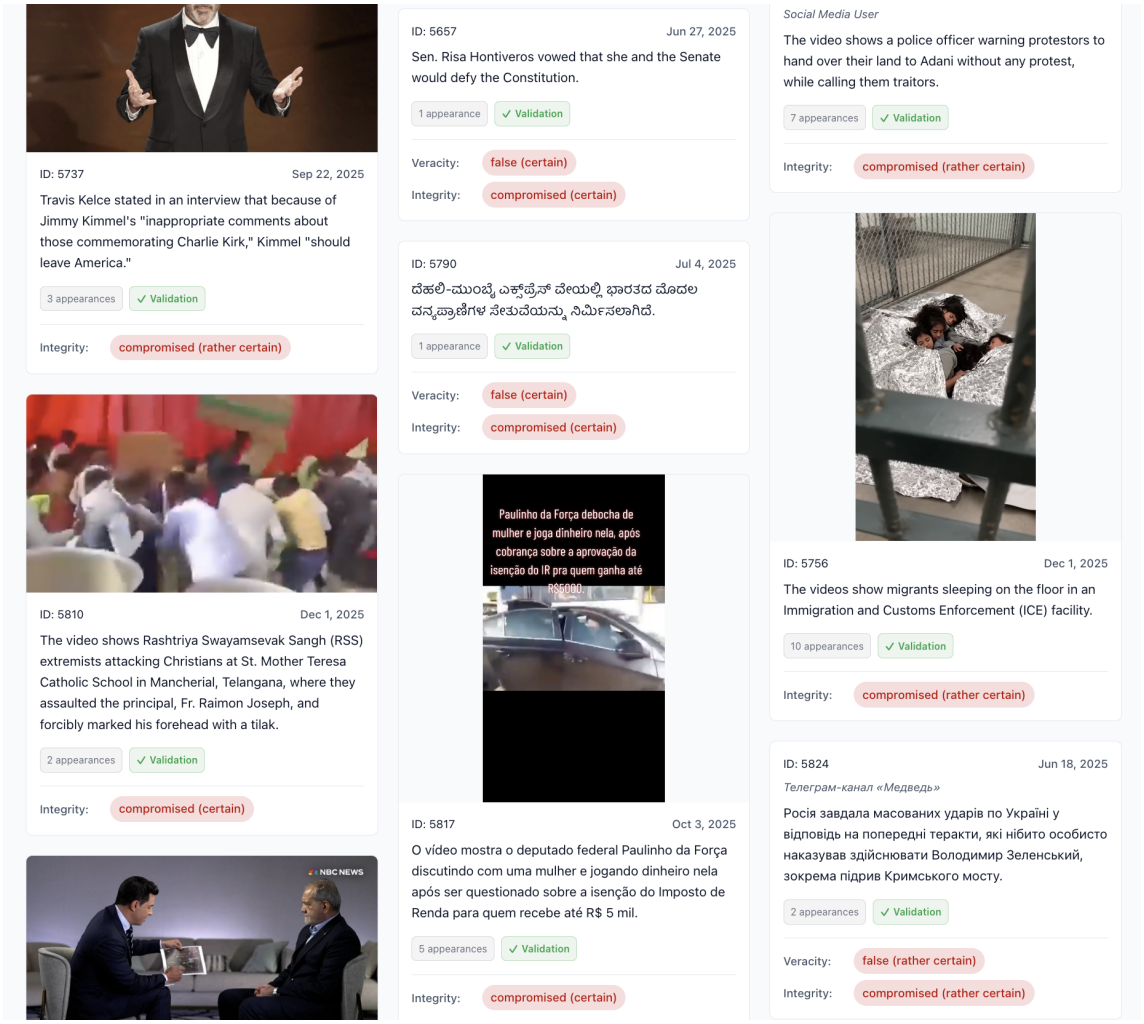
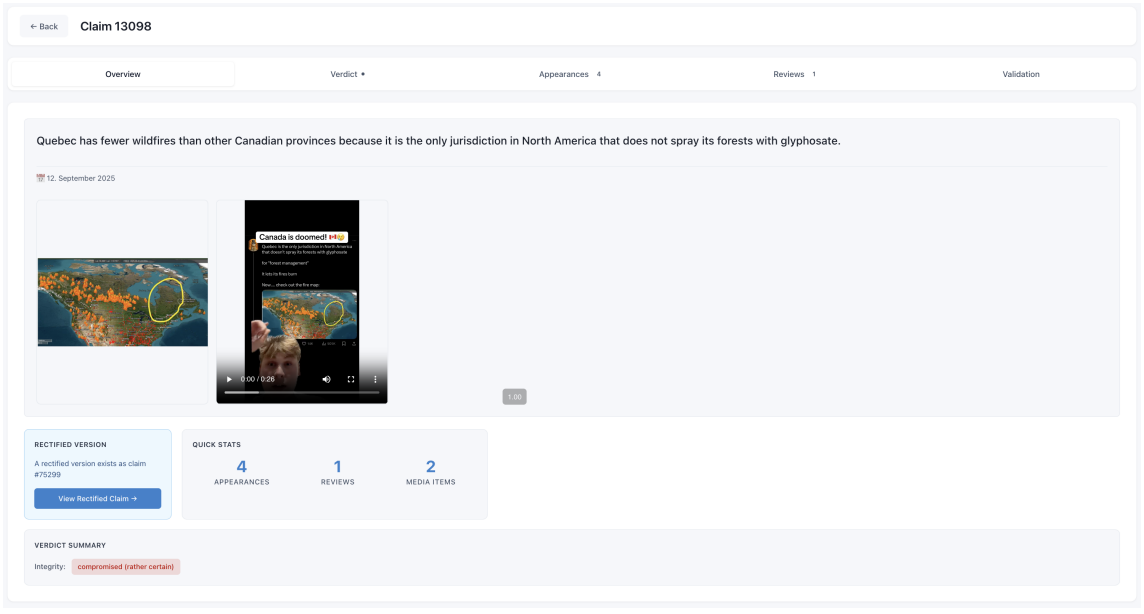


Figure 25: Confusion matrices comparing human annotations with VERITAS' automated annotations.



(a) Screenshot of the Claim Browser UI



(b) Screenshot of the detail-page of a claim from the Claim Browser UI

Figure 26: Screenshots from the Claim Browser UI showing the claim overview and details for a specific claim.