Take It All: Ensemble Retrieval for Multimodal Evidence Aggregation

Max Upravitelev[1,2], Veronika Solopova[1,2], Premtim Sahitaj[1,2], Ariana Sahitaj[1,2], Charlott Jakob[1,2], Sebastian Möller[1,2], and Vera Schmitt[1,2,3,4]

[1]Technische Universität Berlin

[2]German Research Center for Artificial Intelligence (DFKI)

[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data

[4]Centre for European Research in Trusted AI (CERTAIN)

Correspondence: max.upravitelev@tu-berlin.de

Abstract

Multimodal fact checking has become increasingly important due to the predominance of visual content on social media platforms, where images are frequently used to enhance the credibility and spread of misleading claims, while generated images become more prevalent and realistic as generative models advance. Incorporating visual information, however, substantially increases computational costs, raising critical efficiency concerns for practical deployment. In this study, we propose and evaluate the ADA-AGGR (ensemble retrievAl for multimoDAl evidence AGGRegation) pipeline, which achieved the second place on both the dev and test leaderboards of the FEVER 9/AVerImaTeC shared task. However, long runtimes per claim highlight challenges regarding efficiency concerns when designing multimodal claim verification pipelines. We therefore run extensive ablation studies and configuration analyses to identify possible performance–runtime improvements. Our experiments show that substantial efficiency gains are possible without significant loss in verification quality. For instance, we reduced the average runtime by up to 6.28× while maintaining comparable performance across evaluation metrics by aggressively downsampling input images processed by visual language models. Overall, our results highlight that careful design choices are crucial for building scalable and resource-efficient multimodal fact-checking systems suitable for real-world deployment.

1 Introduction

Misinformation on the web increasingly appears in multimodal formats that combine textual claims with images or other media (Akhtar et al., 2023; Cao et al., 2025). Such content is particularly common on social media, where attaching an image to a claim can increase its perceived credibility and potential for the amplification of misleading narratives (Hameleers et al., 2020). This development raises additional challenges for fact-checking: a tweet, meme, or blog post can pair a false textual statement with a compelling image, which may make the overall claim more convincing and faster to spread than text alone. Examples include miscaptioned images, generated deepfakes, and fabricated infographics, which require verification approaches that analyze both textual and visual modalities. In such cases, fact-checkers need to assess extracted information jointly rather than treating either modality in isolation. At the same time, the increasing volume and velocity of online mis- and disinformation far exceed the available capacities of human fact-checkers. Consequently, automated fact-checking (AFC) systems are needed to support human decision-making and scale up verification efforts (Nakov et al., 2021). Such systems should help retrieve relevant evidence, structure the reasoning process, verify claims, and provide intermediate signals, especially in cases where sources are biased or incomplete.

From a practical perspective, AFC systems face two central requirements. First, they must be able to handle multimodal inputs: textual claims accompanied by one or more images, and evidence that can itself be text-only, image-only, or a combination of both. Second, they should operate under limited computational resources and time constraints, achieving accurate and timely verification for both fast-paced newsroom settings and more in-depth investigative journalism workflows. This motivates the use of efficient, open-weight models for AFC systems, enabling realistic deployment beyond heavily resourced laboratory setups. Our work is guided by this practicality: we aim towards the exploration of systems that not only address the challenges of multimodal claim verification, but are also mindful of runtime, resource usage, and deployability in real-world fact-checking environments.

The contributions of this paper are as follows:

Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 91–107
March 28, 2026 ©2026 Association for Computational Linguistics

1. We introduce ADA-AGGR, a pipeline whose Gemini 3 Pro-centered variant achieved second place out of 17 on the AVerImaTeC dev leaderboard and second place on the test leaderboard. Our code is publicly available for reference[1].

2. We extensively evaluate pipeline variants based on open-weight models with the goal of studying trade-offs between retrieval performance and runtime.

3. We show that aggressive visual downsampling is a key efficiency control knob for multimodal fact-checking, with minimal performance loss.

4. We thoroughly analyze potential error sources of different pipeline configurations.

2 Related Work and Preliminaries

Multimodal Retrieval In recent years, approaches targeting retrieval augmented generation (RAG) became a prominent area within the information retrieval (IR) domain. Here, one common goal focuses on the identification and the retrieval of relevant information from documents to augment prompts of Large Language Models (LLMs), a strategy which can be more efficient than including complete documents into prompts directly (Li et al., 2024b). Several approaches have emerged in this area, such as the usage of retriever ensembles comprising information from multiple knowledge sources (Li et al., 2024a; Sanniboina et al., 2024). One challenge in document-based RAG is the multimodality of many documents, which can contain data such as tabular or visual information next to textual data. Hence, the exploration of multimodal RAG approaches emerged, with benchmarks like REAL-MM-RAG (Wasserman et al., 2025) having been released to track progress. In contrast to text-only RAG, multimodal RAG needs to provide for an equivalent of sentence embeddings-based search to create modality-specific or multimodal embeddings for semantic search. Embedding models like ColPali (Faysse et al., 2025) encode visual and textual features within multi-vector embeddings and combine them with the late-interaction mechanism from ColBERT (Khattab and Zaharia, 2020), allowing for the semantic comparison between textual queries and multimodal image patches and enabling state-of-the-art results on multi-modal retrieval benchmarks, as shown in works such as Suri et al. (2025) or Cho et al. (2025).

Multimodal Fact-checking / Claim verification Research on automated fact-checking has traditionally focused on textual claims, for example through benchmark datasets and shared tasks such as AVeriTeC (Schlichtkrull et al., 2024) within the FEVER workshop (Akhtar et al., 2025). These efforts emphasize evidence-based verification pipelines and reproducible, efficient open-weight systems. Beyond text-only settings, multimodal disinformation research shows that pairing text with images can substantially increase the perceived credibility and spread of false claims, motivating fact-checking approaches that incorporate visual evidence (Hameleers et al., 2020). Recent multimodal systems include MIRAGE (Shopnil et al., 2025), a framework for multimodal misinformation detection that combines visual verification, cross-modal consistency checking, and retrieval-augmented fact-checking using large vision-language models (VLM). Complementary work evaluates how VLMs represent multimodal content for fact-checking by training probing classifiers on either joint VLM embeddings or separately fused text and image encoder embeddings which can outperform the intrinsic VLM embeddings for multi-class veracity classification (Cekinel et al., 2025). At retrieval level, analyses of retrieval signals can yield more robust evidence selection in knowledge-intensive tasks (Li et al., 2024a; Sanniboina et al., 2024). Within this landscape, AVerImaTeC situates multimodal fact-checking in a realistic web-based setting where systems must verify image-text claims with both, textual and visual evidence (Cao et al., 2025). Building on this, our work adopts ideas from multimodal verification and retriever ensembles to develop a pipeline which aggregates textual, image-only, and multimodal retrieval components, while targeting efficiency concerns under practical runtime constraints.

3 Methodology

The approach for our system design is inspired by a study of the AVerImaTeC baseline, where each input is assessed to decide which tools to call for retrieving relevant information, such as a reverse image search routine. Within current research, works like Winston and Just (2025) have explored improper tool selection as a source of possible perfor-

[1] https://github.com/XplaiNLP/FEVER9-AVerImaTeC-ADA-AGGR

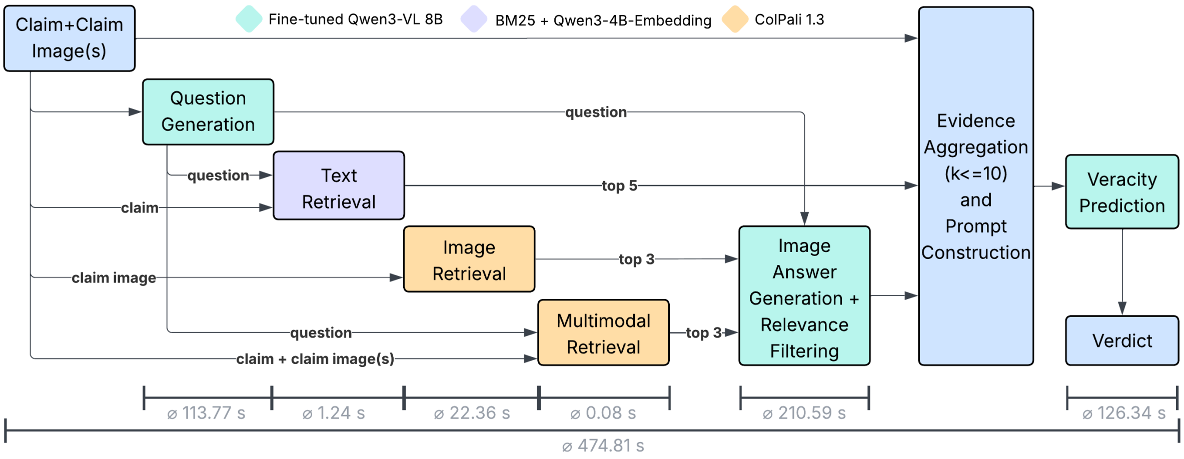

Figure 1: Overview of the proposed pipeline. Note: The documented average durations refer to our configuration based on a fine-tuned Qwen3-VL 8B and not our final system submission, where this model was replaced with Gemini 3 Pro.

mance degradation in RAG systems, while research such as Kalra et al. (2025) have shown strong performance boosts by deploying multiple retrievers and merging their results in a weighted manner. Thus, our central idea is to mitigate error cascades potentially caused by automated tool selection and to cast a wide net instead, using every retrieval method in our pipeline on every input and aggregating the results into a set of 10 evidence pieces (following shared task rules), before passing the aggregated evidence to a final veracity prediction.

Figure 1 presents an overview of our pipeline, including average processing durations of selected components, which are designed as follows:

Question Generation As a first step, we generate a question with a VLM (with the model choice depending on the configuration) based on the input claim together with the claim image(s). This technique is common in recent evidence retrieval pipelines (Schlichtkrull et al., 2024) and essentially equates to query expansion while mirroring the data structure of datasets like AVerImaTeC. We follow the SFEFC approach Upravitelev et al. (2025b) and generate one single question per claim.

Text Retrieval For text-only retrieval, we follow a similar hybrid search approach as in Yoon et al. (2024), where top-k potential candidates are retrieved with sparse retrieval and re-ranked with dense retrieval. For the choice of hyperparameters, we follow SFEFC and retrieve top $k = 1000$ candidates with BM25. Unlike in SFEFC, we increase the chunk size from 4 to 8 due to better results in preliminary experiments and only use the top $k = 5$ dense-retrieval documents as text evidence and reserve the remaining 5 evidence slots for image-based retrieval.

Visual Retrievers Our visual retrieval steps are comprised of two components, both based on returning top $k = 3$ most relevant candidates based on embeddings created with ColPali (Faysse et al., 2025). The k value was determined in preliminary experiments which showed that in most cases, at least one image is filtered out, thus aligning with the shared task rules of collecting 10 evidence pieces per input (to ensure this limit we also truncate the evidence to 10 during the evaluation). The "Image Retrieval" component is similar to reverse image search but uses visual embeddings similar to textual embeddings in dense retrieval (instead of, for example, visual feature extraction), thus returning top k candidates by relevance. The processing time of this component documented in Figure 1 reflects the creation of image embeddings, which are reused in "Multimodal Retrieval". While the first component only retrieves candidates based on image data, the second one also includes the textual claim and the generated question as part of the query. Finally, the candidate sets from both components are merged and de-duplicated before being passed to the next component.

Image Answer Generation Each retrieved image is used together with the generated question

#	Visual Retriever	Text Retriever	Question Generation	Image Answer Generation	Verdict Prediction	Q	E	V	J	Acc	SUP	REF	NEE	CoC	Øs
1	ColPali	Q3-E-4B	Q3-VL 8B FT	Q3-VL 8B FT	Q3-VL 8B FT	0.243	0.267	0.336	0.049	0.711	0.25	0.74	0.33	0.00	474.81
2	ColPali	Q3-E-4B	Q3-VL 8B	Q3-VL 8B	Q3-VL 8B	0.237	0.292	0.367	0.055	0.697	0.75	0.72	0.33	0.00	416.43
3	EvoQwen	Q3-E-4B	Q3-VL 8B FT	Q3-VL 8B FT	Q3-VL 8B FT	0.239	0.287	0.373	0.046	0.743	0.25	0.78	0.33	0.00	601.95
4	ColPali	Q3-E-0.6B	Q3-VL 8B FT	Q3-VL 8B FT	Q3-VL 8B FT	0.233	0.279	0.346	0.046	0.737	0.25	0.77	**0.50**	0.00	419.76
5	ColPali	Q3-E-4b	Gemini 2.5 flash	Gemini 2.5 flash	Gemini 2.5 flash	0.237	0.226	0.251	0.069	0.612	0.25	0.64	0.33	0.00	**159.61**
6	ColPali	Q3-E-4b	Gemini 2.5 pro	Gemini 2.5 pro	Gemini 2.5 pro	**0.294**	0.230	0.307	0.054	0.737	0.50	0.77	0.17	**1.00**	235.44
7	ColPali	Q3-E-4b	Gemini 3 Pro	Gemini 3 Pro	Gemini 3 Pro	0.289	**0.298**	**0.447**	0.050	**0.901**	0.75	**0.94**	0.17	0.00	357.47
8	ColPali	Q3-E-4b	Q3-VL 2B FT	Q3-VL 2B FT	Q3-VL 2B FT	0.179	0.252	0.194	**0.123**	0.434	**1.00**	0.47	0.33	0.00	397.16
9	ColPali	Q3-E-4b	Q3-VL 2B	Q3-VL 2B	Q3-VL 2B	0.203	0.276	0.247	0.121	0.487	**1.00**	0.50	0.00	0.00	375.17
10	ColPali	Q3-E-4b	Q3-VL 8B FT	Q3-VL 2B	Q3-VL 8B FT	0.245	0.258	0.294	0.063	0.612	0.50	0.64	0.17	0.00	433.44
11	ColPali	Q3-E-4b	Q3-VL 8B FT	Q3-VL 2B	Q3-VL 2B	0.247	0.270	0.257	0.098	0.520	**1.00**	0.53	0.00	0.00	419.76
12	ColPali	Q3-E-4b	Q3-VL 8B FT	Q3-VL 8B FT	Q3-VL 2B	0.244	0.291	0.372	0.103	0.750	**1.00**	0.78	0.00	0.00	409.55
13	ColPali	Q3-E-4b	Q3-VL 2B	Q3-VL 8B FT	Q3-VL 8B FT	0.218	0.284	0.319	0.065	0.599	0.50	0.62	0.33	0.00	405.71

Table 1: Evaluation results for different ADA-AGGR configurations. Reported metrics are: Question Score (Q), Evidence Score (E), Verdict Accuracy (V), Accuracy (Acc), Justification Score (J), Runtime (Øs) as well as accuracy results per label: Supported (SUP), Refuted (REF), Not Enough Information (NEE) and Conflicting Evidence/Cherrypicking (CoC); "ft" is short for fine-tuned; "Q3" is short for Qwen3. "E" for Embedding. Highest scored values are bold; underlined values highlight the best results with OS models.

and the claim to prompt a VLM to generate an answer to the question. The prompt within this VLM call also includes a binary filtering criterion: the option for the model to mark the image as irrelevant to the question, in which case it is filtered out before passing the final set of retrieved images and answers to evidence aggregation.

Veracity Prediction After a prompt is constructed with the information from the previous steps, it is passed to a VLM for a final justification generation and veracity prediction. To better guide the model, we include possible reasons for justifying refutation, which we collected from the training data labels. Our preliminary tests showed better results on label accuracy (an ablation of this prompt is evaluated in configuration 7, Table 1). As for the previous steps where a VLM is called, we document our prompts in Appendix B.

Fine-tuning strategy For our experiments, we fine-tuned Qwen3-VL 8B and 2B models (chosen based on good benchmark results (Bai et al., 2025)) on the full training data for a single epoch to counter the risk of overfitting the model toward the REF class in the dataset, which accounts for 95.3% of all labels in the training split. Instead of fine-tuning separate models for individual components (e.g., question generation vs. veracity prediction), we fine-tune a single model jointly for all components, following the intuition that training on full samples with data corresponding to all subtasks might mitigate the class imbalance and improve generalization. The details of our fine-tuning strategy are documented in Appendix C.

4 Evaluation

Overview The goal of our evaluation is to assess the influence of the pipeline components with regard to performance and runtime. Table 1 documents our results on experiments where the pipeline itself is unchanged, but different models are used within the components. For generative tasks, we compare the influence of different-sized Qwen3-VL open-weight models, as well as results based on the proprietary models from the Gemini 2.5 family (Comanici et al., 2025) and Gemini 3 Pro (Google DeepMind, 2025). Furthermore, we explore the influence of models used for textual and multimodal retrieval, deploying different variants of Qwen3-Embedding models as well as ColPali 1.3 (Faysse et al., 2025) as our main model for multimodal retrieval and "EvoQwen2.5-VL-Retriever-3B-v1"[2] for comparison, chosen due to its higher ranking on the public leaderboard[3] of the ViDoRe (Macé et al., 2025) benchmark.

Table 4 documents several ablation studies where different components were removed from the pipeline. Furthermore, we evaluate the influence of reducing image sizes (by downsampling with Lanczos interpolation using factors of 2, 4, and 8) before passing them to the Qwen3-VL 8B model, thus exploring the trade-off between retrieval performance and runtime by reducing the available visual information.

[2] https://huggingface.co/ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-3B-v1

[3] https://huggingface.co/spaces/vidore/vidore-leaderboard

Metrics The metrics selected in Cao et al. (2025) to evaluate proposed approaches follow different LLM-as-a-judge strategies:

- Question Score (Q-score): Generated questions are scored by an LLM in comparison to reference questions from the gold data;
- Evidence Score (E-score): Similar to the Q-Score, collected evidence texts are judged against human annotated ones;
- Verdict Accuracy (V-score): Focuses on the accuracy of predicted labels, while only counting scores if the E-score is over a λ threshold (0.3);
- Justification Score (J-Score): Also similar to the Q-Score, but involves LLMs judging the generated justifications against references in the gold data while only counting scores if the E-score is above λ.

We use an AVerImaTeC scorer[4] with the model "gemini-2.0-flash-001" selected and all evidence sets truncated to 10 elements, following the shared task rules. To counter possible variance and inconsistencies between metrics generated by LLM-as-a-judge approaches, as explored in works such as (Haldar and Hockenmaier, 2025), we collect and average all metrics results over 3 runs. Additionally, we report a simple unconditional label accuracy metric, accuracy per label and the average processing runtime per claim.

Experimental Setup All experiments were conducted on a machine with an AMD EPYC 9254 CPU and an NVIDIA H100 80GB GPU.

4.1 Results

Model Choices The results of Configuration (C) #1 and #2 in Table 1 highlight the importance of the AVerImaTeC metrics: While our fine-tuning strategy did improve the result on a simple label accuracy metric as well as on the Q-Score, all other metrics yielded slightly worse results. A similar behavior can be seen in the configurations with smaller models, C#8 and C#9. Regarding retrieval model choices, C#3 documents an increase in most performance metrics compared to C#1, while also increasing the runtime by over 125 seconds per claim. Surprisingly, deploying a smaller embedding model from the same model family also improved performance metrics, while slightly adding 3 seconds to the runtime. Configurations #5-#7 compare the performance of the same pipeline, but with Gemini models deployed instead of Qwen3-VL variants. C#5 and C#6 underperform against Qwen3-VL 8B variants, while Gemini 3 Pro enabled the best scores on most metrics. Therefore, this model was chosen for our final submission.

id	Q	E	V	J
HUMANE	0.922	0.583	0.644	0.542
ADA-AGGR-gemini3pro	0.353	0.386	0.453	0.372
AIC CTU	0.822	0.346	0.370	0.303
amntq	0.652	0.291	0.302	0.132
ADA-AGGR Q3-VL	0.341	0.290	0.290	0.201
NP10t	0.576	0.225	0.289	0.167
tt	0.731	0.246	0.282	0.262
ADA-AGGR Q3-VL FT	0.355	0.279	0.269	0.183
colabnguyen082	0.562	0.170	0.236	0.130
XxP	0.372	0.222	0.230	0.175
fv	0.429	0.161	0.164	0.117
fv	0.193	0.140	0.144	0.101

Table 2: Results on the AVerImaTeC Dev Leaderboard. Reported metrics are: Question Score (Q), Evidence Score (E), Verdict Accuracy (V), Justification Score (J), sorted by the V Score and documenting the top 10 results (out of 17 in total) together with additional ADA-AGGR configurations for comparison (cursive).

id	Q	E	V	J
HUMANE	0.890	0.536	0.546	0.556
ADA-AGGR	0.370	0.463	0.537	0.433
AIC CTU	0.807	0.325	0.347	0.304
XxP	0.390	0.270	0.256	0.198
teamName	0.662	0.229	0.256	0.216
REVEAL	0.632	0.277	0.236	0.135
fv	0.289	0.163	0.159	0.131
Baseline	0.555	0.171	0.114	0.132

Table 3: Complete results on the AVerImaTeC Test Leaderboard. Reported metrics are: Question Score (Q), Evidence Score (E), Verdict Accuracy (V), Justification Score (J), sorted by the V Score.

As expected, our studies with smaller models show decreased runtimes, which, for the most part, are accompanied by lower performance scores. However, the smaller Qwen3-VL 2B variants were also the only models to predict all 4 SUP labels correctly. Notably, C#12 showcases a decrease in runtime while outperforming many performance metrics of C#1 and C#2.

[4] https://github.com/abril4416/AVerImaTec_Shared_Task/blob/main/prepare_submission/eval_offline.py

Ablation Studies Configurations #1-#8 in Table 4 showcase the influence of the pipeline components on runtime and performance. In most cases, the ablations led to decreased runtimes as well as lower performance metrics, to varying degrees. Ablating our image retrieval component had the smallest influence on performance metrics, while significantly decreasing the runtime. At the same time, ablating the other retrievers in C#3 and C#4 led to sharp performance drops.

Configurations C#9-C#13 target the reduction of image data, which needs to be handled. The reduction of top-k to a single candidate in C#9 and C#10 yielded competitive performance results against our main configuration, while significantly lowering the runtime. Similarly, the reduction of all image sizes before passing them to VLMs remained on a similar level as most of the results of our main configuration, while substantially cutting the runtime, e.g., by a factor of 6.28× from 474.81 seconds to 75.54 seconds per claim when images are shrunk to one-eighth of their size in C#13.

Leaderboard Results Leaderboard performance confirms that the identified efficiency trade-offs do not merely hold in controlled ablations but also translate to competitive shared-task performance. In Table 2 we document the top 10 results from the official AVerImaTeC Dev Leaderboard[5] extended with two ADA-AGGR configurations for comparison. Our pipeline configuration based on Gemini 3 Pro yielded the highest scores within our own experiments on most metrics and hence was chosen for the final submission, which resulted in second place out of 17 submissions. While our pipeline based on Qwen3-VL 8B yielded competitive performance in the middle ranks, it also indicates the superiority of the non-fine-tuned variant on the official metrics. In Table 3 we document all results from the official AVerImaTeC Test Leaderboard[6], where our pipeline based on Gemini 3 Pro yielded second place.

5 Error Analysis

Predicting SUP Labels From the 4 claims labeled with SUP in the dev split of the AVerImaTeC dataset, 3 are image-and-text-related (ids: 33, 68, 97), while 1 is image-related only (id: 151). Our fine-tuned Qwen3-VL configuration predicted one SUP sample correctly (id: 151), indicating that our fine-tuning strategy did not mitigate the class imbalance of the dataset toward the REF label but instead could have biased the model toward it. This indication is supported by the results of C#1 and C#2, where the non-fine-tuned Qwen3-VL outperformed our fine-tuned version on the SUP label. At the same time, the same pipeline with the smaller and non-fine-tuned Qwen3-VL-2B deployed for the final veracity prediction step predicted all 4 SUP labels correctly. Notably, these predictions were generated based on the same aggregated evidence, with the only change in this pipeline being the smaller model choice in the last step. However,

[5] https://huggingface.co/spaces/FEVER-IT/AVerImaTeC

[6] https://huggingface.co/spaces/FEVER-IT/AVerImaTeCTest

#	Strategy	Q	E	V	J	Acc	Ø s
1	w/o Question Generation	-	0.249	0.342	0.021	0.671	336.95
2	w/o Image Retrieval	0.240	0.270	0.325	0.046	0.724	265.74
3	w/o Text Retrieval	0.239	0.137	0.171	0.033	0.586	435.62
4	w/o Mulitimodal Retrieval	0.247	0.090	0.118	0.027	0.441	290.84
5	wo/ Filtering, but w/ Answer Generation	0.237	0.266	0.314	0.031	0.697	561.98
6	w/o Filtering	0.243	0.277	0.334	0.045	0.658	300.82
7	w/o Answer Generation	**0.248**	0.263	0.314	0.053	0.658	300.82
8	w/o Refutation Reasons Prompt	0.241	0.274	0.325	0.067	0.665	464.49
9	w/ Image Retrieval top k=1	0.212	0.269	0.337	0.037	**0.737**	305.76
10	w/ Image, Multimodal Retrieval top k=1	0.226	**0.288**	**0.351**	**0.070**	0.697	198.51
11	shrink images by factor 2	**0.248**	0.266	0.334	0.053	0.684	142.99
12	shrink images by factor 4	0.244	0.254	0.309	0.051	0.704	85.81
13	shrink images by factor 8	**0.248**	0.273	0.331	0.041	0.645	**75.54**

Table 4: Results of ablation and other reduction studies based on our final fine-tuned ADA-AGGR configuration. Reported metrics are: Question Score (Q), Evidence Score (E), Verdict Accuracy (V), Accuracy (Acc), Justification Score (J), Runtime (Øs) as well as accuracy results per label: Supported (SUP), Refuted (REF), Not Enough Evidence (NEE) and Conflicting Evidence/Cherrypicking (CoC). Highest scored values are bold.

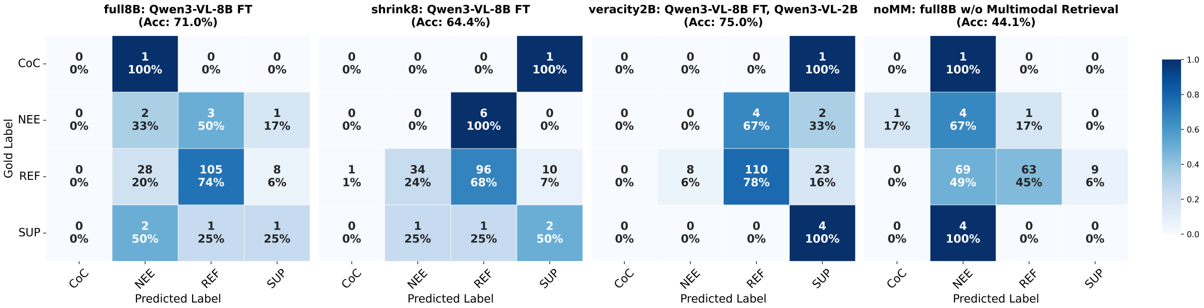

Figure 2: Confusion matrices of selected configurations.

the confusion matrices in Figure 2 display this configuration generally predicting more SUP labels than the others. The results from the configuration without multimodal retrieval show that no SUP labels were identified correctly, indicating that the retrieved textual information was not enough to predict a correct label in the image-and-text-related cases. Furthermore, this configuration also showcases the importance of multimodal evidence by predicting 69 NEE labels while only 6 are found in the dev set.

Caption keywords, helpfulness and confusion For each example in the 4 configurations in Figure 3, we used an LLM `gpt-4o-mini` to annotate which modality was helpful or confusing for deciding the correct gold label. We then normalized the keywords (lowercasing, stripping punctuation) and removed very generic terms (e.g., image, text, evidence, relevant). Figure 3 shows LLM-based helpfulness ratings for each modality across the four configurations. In all cases, retrieved evidence texts receive the highest net helpfulness scores, indicating that textual evidence is the primary driver of correct veracity decisions. Evidence images are on average helpful but exhibit much higher variance and frequent negative outliers, confirming that images are a double-edged sword: they often support the fact-check but can also be severely misleading when reused or symbolic.

The no-multimodal `noMM` configuration displays the lowest medians and widest spreads, especially for evidence images, consistent with the intuition that the system is exposed to less targeted and more ambiguous evidence. Fine-tuning the 8B model (`full8B`) shifts all modalities upwards, particularly evidence images, suggesting that the full system better aligns retrieved evidence with the gold verdict. The smaller `veracity2B` variant and the `shrink-8` model broadly mirror this pattern, with `shrink-8` only slightly reducing the helpfulness of evidence images. This aligns with the quantitative results: aggressive downsampling of images yields substantial efficiency gains while preserving most of their effective contribution to veracity prediction.

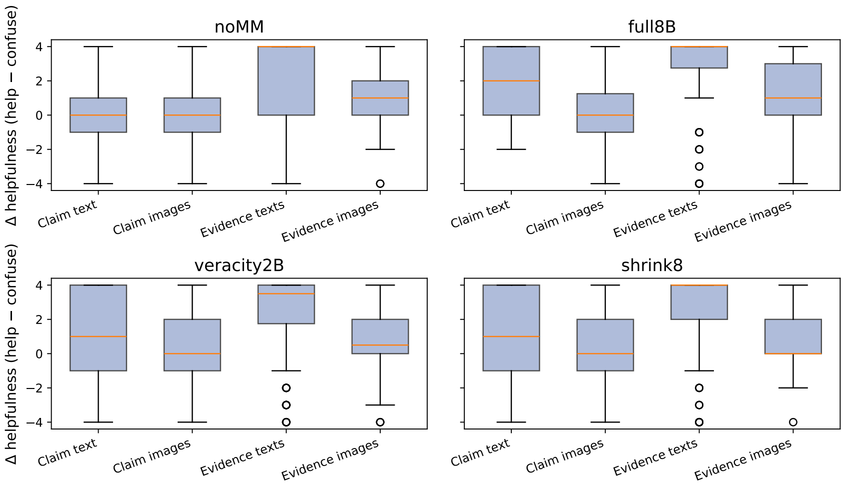

Figure 3: Net helpfulness per modality across configurations.

LLM-based modality ratings also show that textual evidence is the primary driver of correct decisions: the net help–confuse score for retrieved texts is highest, while claim text and claim images are close to neutral. Retrieved images tend to be helpful overall but exhibit substantial variance, with some examples rated as highly misleading (e.g., reused or generated visuals). This supports our design choice that strong text retrieval is essential for factual verdicts, while images play a secondary but sometimes critical role. Additional error analysis is discussed in Appendix 6.

6 Discussion and Conclusion

Our studies explored trade-offs between retrieval performance and runtime of multimodal claim verification pipelines. We showed that in some cases, applying smaller models in specific components does not necessarily cause a performance drop while leading to a decrease in runtime at the same time. Furthermore, the importance of each component has to be taken into account, as shown in our ablation studies. In some cases, an ablation can lead to a significant decrease in runtime while remaining competitive against full pipelines. However, the most substantial decrease here was achieved by targeting the visual information passed to a VLM by aggressively shrinking images by a factor of up to 8 and thus decreasing runtimes by up to a factor of 6.28× while preserving comparable performance.

Limitations

While our fastest configuration was able to yield competitive results at around 75 seconds per claim, this runtime can be considered too long for real-life deployments. One direction of improvement concerns the deployment of VLM models: We worked with unquantized versions of the models, without batching the VLM calls and using the standard transformers library (Wolf et al., 2020) to ensure a better comparison between the configurations. To evaluate the runtime in more common real-life deployments, parallelization techniques such as batching and the use of inference engines like vllm (Kwon et al., 2023) with PagedAttention should be considered. Another direction could be the assessment of which decoder-only VLM/LLM calls might be replaced with other smaller, fine-tuned models. To further mitigate the higher hardware requirements of multimodal data, more visual information reduction techniques like reducing color channels can be explored in future work. As shown in (Yang and Rocha, 2024), fine-tuned encoder–decoder models can perform similarly to larger LLMs on datasets like AVeriTeC, while works such as (Upravitelev et al., 2025a) have demonstrated comparable performance for pipelines based on encoder-only models at a fraction of the runtime of larger models in zero-shot settings on fact-checking datasets such as ClimateCheck (Abu Ahmad et al., 2025). This direction could be explored further in the context of multimodal claim verification. Besides efficiency concerns, our system is prone to performance degradation over time, a phenomenon explored in related fields like automated propaganda detection in recent works such as (Solopova et al., 2025), highlighting the need for systems to account for the chronological and geographical evolution of language.

Ethics Statement

This work proposes a system for automated multimodal claim verification. Automated veracity assessments can be incorrect, biased, or incomplete. Therefore, our system is intended only as a decision-support tool and should not replace human fact-checkers or make final judgments without human oversight. Moreover, our pipeline currently provides only generated justifications and lacks additional explainability features, while studies such as Schmitt et al. (2025) emphasize the importance of explainability features to mitigate users' overreliance on opaque model predictions.

Acknowledgments

The work on this paper is performed in the scope of the projects "VeraXtract" (16IS24066) and "newspolygraph" (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 42–56, Vienna, Austria. Association for Computational Linguistics.

Mubashara Akhtar, Rami Aly, Yulong Chen, Zhenyun Deng, Michael Schlichtkrull, Chenxi Whitehouse, and Andreas Vlachos. 2025. The 2nd automated verification of textual claims (AVeriTeC) shared task: Open-weights, reproducible and efficient systems. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 201–223, Vienna, Austria. Association for Computational Linguistics.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2025. M3docvqa: Multi-modal multi-page multi-document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6178–6188.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Google DeepMind. 2025. Gemini 3. Google DeepMind Model Release. Accessed: 2025-11-30.

Rajarshi Haldar and Julia Hockenmaier. 2025. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, Suzhou, China. Association for Computational Linguistics.

Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political communication*, 37(2):281–301.

Jushaan Singh Kalra, Xinran Zhao, To Eun Kim, Fengyu Cai, Fernando Diaz, and Tongshuang Wu. 2025. MoR: Better handling diverse queries with a mixture of sparse, dense, and human retrievers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11982–12001, Suzhou, China. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mingda Li, Xinyu Li, Yifan Chen, Wenfeng Xuan, and Weinan Zhang. 2024a. Unraveling and mitigating retriever inconsistencies in retrieval-augmented large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4833–4850, Bangkok, Thailand. Association for Computational Linguistics.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark v2: Raising the bar for visual retrieval. *Preprint*, arXiv:2505.17166.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino.

2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4551–4558, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Saikrishna Sanniboina, Shiv Trivedi, and Sreenidhi Vijayaraghavan. 2024. Lore: Logit-ranked retriever ensemble for enhancing open-domain question answering. *Preprint*, arXiv:2410.10042.

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. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Vera Schmitt, Isabel Bezzaoui, Charlott Jakob, Premtim Sahitaj, Qianli Wang, Arthur Hilbert, Max Upravitelev, Jonas Fegert, Sebastian Möller, and Veronika Solopova. 2025. Beyond transparency: Evaluating explainability in ai-supported fact-checking. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, MAD' 25, page 63–72, New York, NY, USA. Association for Computing Machinery.

Mir Nafis Sharear Shopnil, Sharad Duwal, Abhishek Tyagi, and Adiba Mahbub Proma. 2025. Mirage: Agentic framework for multimodal misinformation detection with web-grounded reasoning. *Preprint*, arXiv:2510.17590.

Veronika Solopova, Robert Nickel, and Dorothea Kolossa. 2025. *Lagging Behind: Challenges of Adapting Automated Propaganda Detection to the Chronological and Geographic Evolution of Language*, chapter Chapter 9.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6088–6109, Albuquerque, New Mexico. Association for Computational Linguistics.

Max Upravitelev, Nicolau Duran-Silva, Christian Woerle, Giuseppe Guarino, Salar Mohtaj, Jing Yang, Veronika Solopova, and Vera Schmitt. 2025a. Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 281–287, Vienna, Austria. Association for Computational Linguistics.

Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. 2025b. Exploring semantic filtering heuristics for efficient claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 229–237, Vienna, Austria. Association for Computational Linguistics.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31660–31683, Vienna, Austria. Association for Computational Linguistics.

Cailin Winston and René Just. 2025. A taxonomy of failures in tool-augmented llms. In *2025 IEEE/ACM International Conference on Automation of Software Test (AST)*, pages 125–135.

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

Jing Yang and Anderson Rocha. 2024. Take it easy: Label-adaptive self-rationalization for fact verification and explanation generation. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Error Analysis II: Part-of-Speech and Named Entity Recognition

We characterised the linguistic content of each example separately for four channels: claim text, claim image captions, evidence texts, and evidence image captions. The captions were generated by prompting a Qwen3-VL 8B model. For every channel, we concatenated all available snippets and ran the English `en_core_web_sm` pipeline from spaCy to obtain: (i) total number of tokens and nouns, (ii) number of named entities (spaCy NER), (iii) number of numeric expressions (tokens with `like_num=True` or entity types such as CARDINAL, DATE, MONEY, etc.), (iv) counts of abstract vs. concrete nouns.

To distinguish abstract and concrete nouns, we first collected the lemma vocabulary of all nouns across the dataset and asked an LLM (`gpt-4o-mini`) to label each lemma as ABSTRACT, CONCRETE or OTHER. We then mapped each noun token to this label and computed, per channel and per example, the proportion of abstract vs. concrete nouns among all nouns. The claim itself averages only ≈17 tokens and about 2.5 named entities, which matches the intuition that social media claims tend to be short, lossy summaries of an alleged event. texts and especially evidence captions are much longer (≈128 and ≈434 tokens on average) and contain many more entities and numbers. Evidence captions, which concatenate descriptions of all retrieved images, are the most "structured" channel with ≈17 entities and ≈8.6 numeric mentions per example.

Interestingly, correct classifications are slightly more abstract (contain a bit more abstract nouns). Claims and retrieved evidences of correct predictions, also contain less numbers and named entities, suggesting that the model performs better where it might be able able to use common sense or not have to analyse exact numbers.

```
================full8B================
- Evidence TEXT: misclassified (top 20)
```

```
[('image', 288), ('said', 90), ('shows', 90),
('flag', 80), ('claim', 75), ('people', 58),
('photo', 57), ('evidence', 50), ('police',
46), ('u.s.', 45), ('yes', 44), ('provide',
43), ('law', 39), ('missing', 37), ('relevant',
36), ('', 36), ('national', 35), ('woman', 35),
('consistent', 34), ('new', 33), ('thailand',
32), ('pride', 32), ('code', 31), ('ukrainian',
30), ('news', 30), ('states', 30),
('information', 29), ('government', 29),
('years', 29), ('person', 28), ('violence',
28), ('flags', 28), ('white', 27), ('waiting',
27), ('ukraine', 26), ('military', 25),
('house', 24), ('lists', 24), ('rock', 24),
('', 24), ('president', 23), ('research', 23),
('day', 22), ('year', 22), ('boat', 22),
('protest', 21), ('confirm', 21), ('hogy', 21),
('official', 20), ('china', 20)]
```

```
- Evidence TEXT: correct (top 20)
```

```
[('image', 863), ('shows', 302), ('said', 271),
('claim', 237), ('photo', 229), ('police',
219), ('people', 217), ('modi', 147), ('woman',
129), ('india', 125), ('evidence', 124),
('new', 112), ('hospital', 106), ('french',
105), ('', 102), ('man', 102), ('national',
97), ('que', 95), ('raja', 95), ('state', 94),
('time', 91), ('video', 91), ('yes', 89),
('year', 86), ('june', 83), ('years', 83),
('provide', 83), ('world', 82), ('area', 81),
('confirm', 79), ('news', 79), ('media', 77),
('person', 77), ('manipur', 76), ('minister',
75), ('appears', 75), ('false', 74), ('taken',
74), ('work', 74), ('old', 72), ('based', 72),
('visible', 72), ('bjp', 71), ('city', 70),
('social', 70), ('real', 69), ('visit', 67),
('viral', 67), ('like', 67), ('information',
67)]
```

```
- Evidence IMAGE CAPTIONS: misclassified
(top 20)
```

```
[('image', 553), ('text', 292), ('white', 235),
('notice', 215), ('flag', 150), ('visible',
131), ('photograph', 128), ('appears', 118),
('man', 100), ('likely', 96), ('change', 91),
('shirt', 86), ('background', 82), ('right',
81), ('holding', 80), ('blue', 76),
('different', 76), ('house', 74), ('scene',
70), ('left', 69), ('rainbow', 66), ('black',
64), ('bank', 63), ('photo', 61), ('body', 60),
('wearing', 59), ('red', 59), ('includes', 59),
('shows', 58), ('dark', 58), ('possibly', 56),
```

channel	mean_tokens	mean_nouns	mean_ents	mean_numbers	mean_abs_ratio	mean_con_ratio
0 Claim Text	16.82	7.16	2.52	0.95	0.15	0.50
1 Claim Captions	190.41	59.18	5.93	3.32	0.20	0.58
2 Evidence Texts	128.45	42.23	8.54	3.51	0.20	0.60
3 Evidence Captions	433.76	141.94	16.75	8.57	0.21	0.58

Table 5: Linguistic channel-level summary (tokens, nouns, entities, numeric mentions, abstract/concrete ratios).

```
(’person’, 56), (’green’, 51), (’large’, 48),
(’suitcase’, 48), (’customer’, 48), (’setting’,
46), (’inside’, 46), (’article’, 44),
(’display’, 42), (’location’, 42), (’pride’,
42), (’including’, 40), (’branch’, 40),
(’taken’, 39), (’light’, 38), (’cash’, 38),
(’limit’, 38), (’temporary’, 38),
(’withdrawal’, 37)]
```

− Evidence IMAGE CAPTIONS: correct (top 20)

```
[(’image’, 1534), (’text’, 772), (’visible’,
407), (’appears’, 346), (’white’, 330),
(’background’, 269), (’red’, 254), (’likely’,
242), (’right’, 235), (’left’, 232), (’blue’,
230), (’dark’, 228), (’large’, 219), (’map’,
217), (’photograph’, 214), (’wearing’, 203),
(’scene’, 202), (’photo’, 180), (’man’, 180),
(’black’, 170), (’showing’, 156), (’woman’,
148), (’shows’, 148), (’light’, 144),
(’includes’, 144), (’person’, 142), (’post’,
124), (’setting’, 116), (’possibly’, 116),
(’green’, 116), (’police’, 116), (’including’,
112), (’shirt’, 105), (’people’, 104),
(’yellow’, 103), (’street’, 103), (’real’,
101), (’graphic’, 96), (’news’, 95),
(’protest’, 95), (’road’, 92), (’suggesting’,
91), (’sky’, 91), (’event’, 89), (’trees’, 89),
(’near’, 87), (’claim’, 85), (’area’, 85),
(’composite’, 85), (’standing’, 83)]
```

================noMM================

− Evidence TEXT: misclassified (top 20)

```
[(’image’, 361), (’shows’, 118), (’flag’, 77),
(’claim’, 75), (’evidence’, 63), (’relevant’,
61), (’yes’, 55), (’provide’, 44), (’confirm’,
40), (’consistent’, 37), (’police’, 36),
(’photo’, 33), (’appears’, 30), (’visible’,
30), (’person’, 29), (’woman’, 28), (’context’,
26), (’visual’, 26), (’pride’, 25), (’states’,
24), (’u.s.’, 24), (’white’, 23), (’modi’, 23),
(’man’, 22), (’verified’, 22), (’code’, 21),
(’june’, 19), (’flags’, 19), (’lists’, 19),
(’solely’, 18), (’waiting’, 18), (’based’, 17),
(’poster’, 17), (’protesters’, 16), (’wearing’,
15), (’accident’, 15), (’text’, 15),
(’verification’, 15), (’muslim’, 15),
(’specific’, 14), (’confirmed’, 14), (’french’,
14), (’hand’, 14), (’images’, 14), (’rob’, 14),
(’displayed’, 14), (’verify’, 14), (’real’,
14), (’appear’, 13), (’verifiable’, 13)]
```

− Evidence TEXT: correct (top 20)

```
[(’image’, 337), (’shows’, 112), (’raja’, 90),
(’claim’, 69), (’photo’, 67), (’ravi’, 55),
(’varma’, 55), (’contemporary’, 45),
(’evidence’, 38), (’french’, 38), (’riots’,
36), (’confirm’, 34), (’visual’, 33),
(’visible’, 33), (’woman’, 32), (’occurred’,
31), (’appears’, 31), (’depicts’, 31),
(’provide’, 30), (’baby’, 27), (’yes’, 24),
(’based’, 24), (’false’, 23), (’near’, 22),
(’context’, 22), (’shown’, 21), (’information’,
21), (’mosque’, 20), (’depict’, 20),
(’consistent’, 20), (’person’, 20), (’real’,
20), (’man’, 19), (’india’, 19), (’photograph’,
19), (’protest’, 18), (’relevant’, 18),
(’likely’, 17), (’trump’, 17), (’world’, 17),
(’highway’, 17), (’specific’, 16), (’site’,
16), (’train’, 16), (’wearing’, 16), (’solely’,
16), (’police’, 16), (’taken’, 16), (’text’,
15), (’location’, 15)]
```

− Evidence IMAGE CAPTIONS: misclassified (top 20)

```
[(’image’, 621), (’visible’, 247), (’appears’,
237), (’white’, 226), (’poor’, 202),
(’richard’, 201), (’official’, 191), (’poster’,
157), (’background’, 131), (’likely’, 129),
(’blue’, 128), (’text’, 121), (’photograph’,
115), (’left’, 113), (’scene’, 111), (’red’,
107), (’setting’, 104), (’right’, 104),
(’wearing’, 101), (’real’, 101), (’dark’, 100),
(’possibly’, 96), (’police’, 96), (’man’, 89),
(’black’, 87), (’large’, 82), (’green’, 79),
(’public’, 75), (’event’, 73), (’taken’, 72),
(’government’, 69), (’protest’, 69), (’shows’,
67), (’lighting’, 66), (’photo’, 66), (’flag’,
63), (’suggests’, 62), (’holding’, 60),
(’light’, 56), (’satellite’, 56), (’shirt’,
55), (’including’, 55), (’yellow’, 55),
(’exhibition’, 50), (’publication’, 50),
(’hand’, 49), (’outdoor’, 49), (’form’, 49),
(’stock’, 49), (’advertisement’, 48)]
```

− Evidence IMAGE CAPTIONS: correct (top 20)

```
[(’image’, 455), (’map’, 201), (’visible’,
186), (’white’, 160), (’appears’, 151),
(’ocean’, 135), (’photograph’, 130), (’dark’,
125), (’background’, 115), (’scene’, 107),
(’red’, 106), (’text’, 101), (’blue’, 100),
(’right’, 97), (’pacific’, 96), (’likely’, 95),
(’large’, 94), (’real’, 91), (’left’, 86),
(’woman’, 84), (’wearing’, 81), (’man’, 81),
(’light’, 79), (’public’, 76), (’black’, 75),
(’indian’, 66), (’including’, 61), (’green’,
59), (’multiple’, 58), (’data’, 57), (’region’,
57), (’billboard’, 57), (’includes’, 56),
(’modi’, 56), (’setting’, 54), (’people’, 54),
(’shown’, 52), (’shows’, 51), (’street’, 51),
(’protest’, 51), (’road’, 51), (’suggesting’,
49), (’taken’, 49), (’hair’, 48), (’area’, 47),
(’possibly’, 47), (’outside’, 46), (’small’,
44), (’buildings’, 44), (’composite’, 44)]
```

================veracity2B=============

− Evidence TEXT: misclassified (top 20)

```
[(’image’, 308), (’yes’, 164), (’shows’, 133),
(’claim’, 121), (’photo’, 106), (’flag’, 89),
(’said’, 87), (’police’, 82), (’people’, 78),
(’national’, 58), (’vcg’, 56), (’relevant’,
55), (’consistent’, 54), (’china’, 52),
(’woman’, 51), (’video’, 49), (’thailand’, 47),
(’military’, 46), (’law’, 46), (’titanic’, 45),
(’government’, 45), (’pride’, 45), (’rock’,
42), (’manipur’, 41), (’les’, 40), (’news’,
40), (’bridge’, 40), (’white’, 40), (’house’,
40), (’ukrainian’, 39), (’images’, 39),
(’road’, 38), (’’, 37), (’violence’, 37),
(’evidence’, 36), (’area’, 35), (’highway’,
35), (’years’, 35), (’volcano’, 35), (’near’,
34), (’information’, 34), (’west’, 34),
(’viral’, 34), (’missing’, 34), (’province’,
33), (’meta’, 33), (’new’, 32), (’village’,
32), (’label’, 32), (’arrested’, 32)]
```

– Evidence TEXT: correct (top 20)

```
[(’image’, 519), (’said’, 274), (’yes’, 217),
(’people’, 194), (’shows’, 176), (’claim’,
162), (’police’, 151), (’photo’, 127), (’modi’,
113), (’new’, 110), (’india’, 105), (’’, 101),
(’man’, 96), (’hospital’, 93), (’state’, 90),
(’que’, 86), (’time’, 84), (’old’, 82),
(’relevant’, 81), (’years’, 79), (’world’, 78),
(’year’, 78), (’area’, 75), (’june’, 74),
(’woman’, 73), (’train’, 68), (’flag’, 68),
(’city’, 67), (’work’, 64), (’bjp’, 63),
(’consistent’, 62), (’national’, 62), (’like’,
61), (’according’, 60), (’minister’, 59),
(’twitter’, 59), (’media’, 59), (’video’, 58),
(’day’, 58), (’information’, 56), (’visit’,
56), (’baby’, 55), (’news’, 55), (’real’, 54),
(’images’, 54), (’south’, 53), (’protest’, 53),
(’mosque’, 52), (’north’, 51), (’kuki’, 51)]
```

– Evidence IMAGE CAPTIONS: misclassified (top 20)

```
[(’image’, 785), (’text’, 386), (’white’, 318),
(’visible’, 225), (’appears’, 186), (’flag’,
177), (’house’, 176), (’dark’, 151),
(’background’, 143), (’right’, 124),
(’photograph’, 122), (’left’, 115), (’blue’,
111), (’scene’, 110), (’black’, 108), (’fact’,
108), (’man’, 102), (’wearing’, 101), (’bank’,
97), (’including’, 96), (’taken’, 96), (’red’,
94), (’large’, 92), (’likely’, 91),
(’agreement’, 86), (’shows’, 84), (’rainbow’,
79), (’pride’, 77), (’notice’, 76), (’setting’,
75), (’real’, 67), (’possibly’, 66), (’woman’,
64), (’suggesting’, 63), (’sky’, 63),
(’political’, 63), (’green’, 62), (’people’,
62), (’content’, 62), (’shirt’, 60), (’yellow’,
60), (’includes’, 56), (’light’, 54),
(’display’, 53), (’check’, 53), (’person’, 52),
(’police’, 51), (’checked’, 50), (’holding’,
49), (’rock’, 49)]
```

– Evidence IMAGE CAPTIONS: correct (top 20)

```
[(’image’, 1325), (’text’, 666), (’visible’,
346), (’appears’, 306), (’white’, 226),
(’background’, 214), (’likely’, 188),
(’photograph’, 185), (’scene’, 179), (’map’,
176), (’right’, 173), (’left’, 172), (’large’,
172), (’dark’, 168), (’black’, 160), (’red’,
160), (’wearing’, 156), (’blue’, 153),
(’photo’, 137), (’shows’, 134), (’nayi’, 130),
(’jaise’, 129), (’person’, 119), (’man’, 118),
(’light’, 117), (’setting’, 111), (’post’,
109), (’green’, 104), (’screenshot’, 95),
(’people’, 94), (’possibly’, 93), (’infant’,
92), (’street’, 92), (’surface’, 91),
(’building’, 90), (’showing’, 89), (’media’,
89), (’shirt’, 88), (’suggests’, 88), (’woman’,
87), (’includes’, 87), (’real’, 86), (’police’,
85), (’government’, 81), (’features’, 81),
(’social’, 79), (’baby’, 77), (’holding’, 76),
 (’train’, 76), (’small’, 73)]
```

===============shrink8===============

– Evidence TEXT: misclassified (top 20)

```
[(’image’, 305), (’flag’, 112), (’shows’, 104),
(’people’, 101), (’said’, 91), (’police’, 83),
(’claim’, 79), (’photo’, 62), (’u.s.’, 55),
(’manipur’, 54), (’relevant’, 51), (’south’,
51), (’titanic’, 49), (’new’, 49), (’national’,
48), (’law’, 45), (’pride’, 44), (’evidence’,
43), (’violence’, 43), (’woman’, 43), (’yes’,
42), (’state’, 42), (’video’, 37), (’thailand’,
37), (’government’, 35), (’june’, 34),
(’world’, 34), (’code’, 34), (’based’, 33),
(’year’, 33), (’volcano’, 33), (’country’, 32),
(’arrested’, 32), (’information’, 31),
(’submersible’, 31), (’north’, 31), (’provide’,
31), (’mayon’, 31), (’news’, 30), (’ukrainian’,
30), (’text’, 30), (’years’, 30), (’india’,
30), (’american’, 29), (’military’, 29),
(’protest’, 29), (’eruption’, 29), (’missing’,
29), (’confirm’, 28), (’appears’, 28)]
```

– Evidence TEXT: correct (top 20)

```
[(’image’, 591), (’said’, 269), (’shows’, 211),
(’photo’, 179), (’people’, 171), (’police’,
146), (’modi’, 139), (’claim’, 122), (’’, 117),
(’india’, 101), (’que’, 101), (’hospital’,
100), (’raja’, 96), (’man’, 93), (’new’, 88),
(’yes’, 85), (’woman’, 84), (’national’, 78),
(’year’, 70), (’state’, 70), (’evidence’, 69),
(’relevant’, 69), (’bjp’, 68), (’work’, 67),
(’time’, 67), (’taken’, 66), (’media’, 66),
(’visit’, 65), (’news’, 64), (’day’, 64),
(’twitter’, 64), (’years’, 63), (’missing’,
63), (’area’, 63), (’person’, 63), (’appears’,
61), (’baby’, 60), (’old’, 60), (’video’, 60),
(’world’, 59), (’train’, 59), (’near’, 59),
(’minister’, 57), (’body’, 57), (’social’, 57),
(’according’, 57), (’ravi’, 57), (’like’, 56),
(’found’, 56), (’varma’, 56)]
```

– Evidence IMAGE CAPTIONS: misclassified (top 20)

```
[(’image’, 654), (’text’, 358), (’white’, 281),
(’flag’, 188), (’photograph’, 167), (’map’,
161), (’house’, 158), (’visible’, 152),
(’appears’, 141), (’real’, 104), (’blue’, 103),
(’likely’, 97), (’official’, 96),
(’background’, 94), (’dark’, 92), (’scene’,
92), (’election’, 90), (’red’, 89), (’right’,
83), (’large’, 72), (’black’, 71), (’left’,
70), (’wearing’, 67), (’man’, 63), (’pride’,
62), (’flags’, 60), (’public’, 58), (’news’,
58), (’display’, 56), (’setting’, 55),
(’includes’, 55), (’event’, 52), (’including’,
51), (’shows’, 51), (’logo’, 50),
(’recognized’, 50), (’police’, 48), (’light’,
46), (’source’, 45), (’website’, 45),
(’design’, 45), (’rainbow’, 45), (’counties’,
44), (’green’, 43), (’possibly’, 43),
(’credible’, 43), (’colored’, 42), (’staged’,
42), (’sky’, 42), (’shirt’, 42)]
```

– Evidence IMAGE CAPTIONS: correct (top 20)

```
[(’image’, 1489), (’text’, 649), (’appears’,
329), (’visible’, 304), (’white’, 255),
(’official’, 223), (’likely’, 210), (’light’,
201), (’photo’, 196), (’red’, 191),
('background', 187), ('wearing', 186), ('blue',
179), ('media', 177), ('left', 176), ('right',
175), ('post', 175), ('dark', 171), ('large',
159), ('scene', 159), ('photograph', 150),
('man', 137), ('black', 131), ('possibly',
125), ('people', 120), ('woman', 114),
('green', 114), ('includes', 112), ('social',
112), ('government', 110), ('surface', 109),
('shows', 104), ('news', 101), ('images', 100),
('infant', 99), ('india', 95), ('earth', 93),
('including', 91), ('setting', 90), ('traffic',
90), ('skin', 89), ('police', 87), ('real',
86), ('composite', 86), ('suggests', 84),
('yellow', 82), ('showing', 80),
('individuals', 78), ('rough', 77), ('article',
76)]
```

We computed lemmatised unigram frequencies over all retrieved evidence texts and image captions, separately for correctly classified and misclassified claims in each configuration discussed in the error analysis sectionin the main part of this paper (full8B, noMM, veracity2B, shrink8). In the textual evidence, generic narration terms such as image, shows, claim, photo dominate across all systems, as expected from fact-checking articles that paraphrase the claim ("the image shows ..."). More informative are domain-specific clusters. Correct predictions are frequently associated with explicit truth-status cues (false, real, old, taken) and concrete event descriptors (Modi, India, Manipur, riots, hospital, train, mosque). Misclassified examples, in contrast, show a higher concentration of meta-evidence vocabulary (relevant, consistent, provide, verification, visual, solely) and more symbolic or high-drama topics, including flag, pride, violence and rarer disaster narratives (Titanic, volcano, submersible, Mayon).

In the image captions, we again find a layer of generic visual descriptors (white, visible, background, scene, wearing, man, woman). Beyond this, captions in correctly classified examples disproportionately mention contextualising or explanatory visuals such as map, region, earth, street, billboard, composite, screenshot, as well as explicit media framing (news, media, social, post). Misclassified examples, especially in the noMM and shrink8 configurations, are dominated by references to paintings and stock imagery (poster, exhibition, publication, stock, advertisement), small logos and UI elements (logo, website, source, traffic), and symbolic visuals (rainbow, pride, bank, cash).

Entity counts reveal a strong configuration effect but only a weak correctness effect (Figure 4). The no-multimodal variant retrieves evidence with very few named entities, whereas all multimodal configurations operate on entity-rich texts. However, within each configuration, correct and misclassified examples have broadly similar entity distributions: misclassified cases are not simply those with "too few entities". At the same time we see slightly more volatile distribution in misclassified cases overall. For the full8B system, correct examples contain slightly more entities on average, hinting that better retrieval helps, but in the 2B and shrink-8 setups misclassified examples can even be more entity-dense, suggesting that dense, information-rich evidence is not automatically easier to use and may overload smaller models. A very similar picture can be observed in terms of numerical entities in evidences (Figure 5).

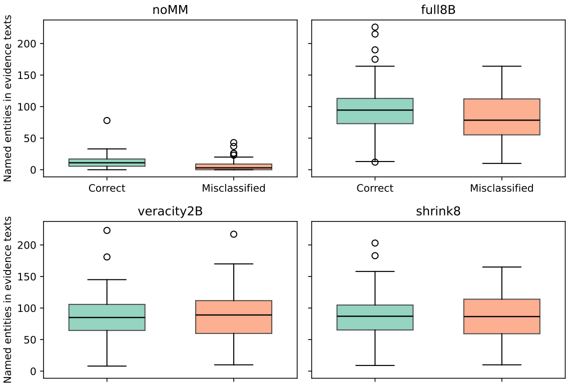

Figure 4: Structured information (entities) vs. correctness across configurations

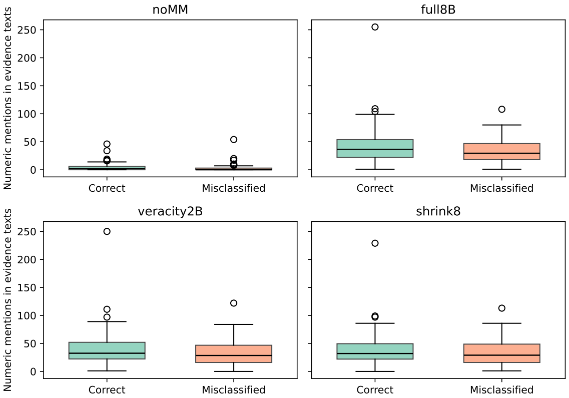

Figure 5: Structured information (numbers) vs. correctness across configurations

Across all configurations, the proportion of abstract nouns in the retrieved evidence texts is very similar for correctly classified and misclassified claims

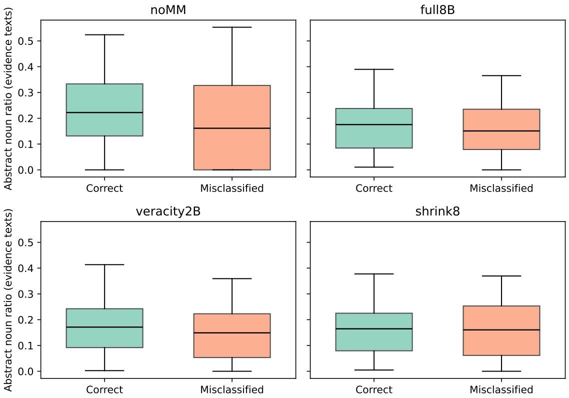

Figure 6: Abstractness of evidence texts vs. correctness across configurations

(Figure 6). Correct and incorrect examples show overlapping distributions, with only small shifts in the median abstractness. This suggests that the overall abstract/concrete balance of the language in the evidence is not a strong driver of errors: hard cases are not simply “too abstract” or “too concrete”, but problematic for more specific reasons (e.g., misleading visuals or complex event structure).

Again, we only see wider boxes (more volatile distributions in misclassified samples), except for `full8B`, the best performing version, where the distribution of correct and misclassified is almost identical, indicating that a better model does improve on abstract concept processing.

B Prompts Collection

B.1 Embedding models

The default model instruction prompt is:

```
f``Instruct: Given a web search query, retrieve
relevant passages that answer the query \n
Query: {query}''
```

B.2 Generative Models

Question Generation

```
"Generate one concise essential verification
question for the following claim.\n"
"Return only the question as plain text.\n"
f"Claim: {claim_text}"
```

The claim images were included in the LLM call.

Image Answer Generation

```
"You are a fact-checking assistant. Your task
is to analyze an image in the context
of a claim and a specific question.\n"
"1. First, determine if the provided image is
relevant for answering the question about the
claim.\n"
"2. If the image is relevant, provide a concise
answer to the question based *only*
on what you can see in the image and the context
from the claim.\n"
"3. If the image is not relevant, state 'Image
is not relevant.' as the answer.\n"
"Respond in a JSON format with two keys:
'relevant' (boolean) and 'answer' (string).\n"
"---\n"
f"Claim: {claim_text}\n"
f"Question: {question}\n"
"---\n"
"Here is the image to analyze:"
```

Veracity Prediction

```
"You are a professional fact-checker. Using
only the provided evidence items
(text and image answers),",
"produce exactly one JSON object and nothing
else with two keys: \"verdict\" and
\"justification\".",
"",
"VERDICT LABELS (choose exactly one):",
" - Supported",
" - Refuted",
" - Not Enough Evidence",
" - Conflicting Evidence/Cherry-picking",
"",
"The justification should be concise and
cite evidence items by their index like
[TEXT_0], [IMAGE_1],",
"or claim images using [CLAIM_IMAGE_0]
notation if you refer to the image that was
provided with the claim.",
"Do NOT hallucinate additional facts; rely
only on the supplied evidence pieces.",
"",
"Possible reasons which can be part of the
justification",
"question_type: Text-related, Image-related,
Metadata-related, Commonsense-related.
\nanswer_type: Abstractive, Extractive,
Unanswerable, Boolean,
Image\nfact_checking_strategies: Written
Evidence, Consultation, Keyword Search,
Numerical Comparison, Reverse Image Search,
Fact-checker Reference, Media Source
Discovery, Image Analysis, Geolocation,
Video
Analysis, Satirical Source Identification,
Audio Analysis\nrefuting_reasons: Misuse
of images, Textual refuted,
Others\nimage_misuse_types:
Out-of-context, Others, ",
"",
"IMPORTANT: Images that are part of the
claim are listed under [CLAIM_IMAGE_i] and
are part of the claim context --",
"they are NOT to be treated as retrieved
evidence. Other images are evidence and are
listed under [IMAGE_i].",
"When actual image files are attached to
this prompt, they are provided in the same
order as shown below:",
"  first: all CLAIM images ([CLAIM_IMAGE_0],
[CLAIM_IMAGE_1], ...),",
"  then: all evidence images ([IMAGE_0],
[IMAGE_1], ...).",
"",
f"Claim: {claim_text}",
f"Question: {question}",
"",
"Claim images (indexed):",
"\n".join(claim_image_prompt_lines) if
claim_image_prompt_lines else "No claim
images provided.",
"",
"Text evidence items (indexed):",
"\n".join(evidence_texts_prompt) if
evidence_texts_prompt else "No text evidence
provided.",
"",
"Image evidence items (indexed):",
"\n".join(evidence_image_prompt_lines) if
evidence_image_prompt_lines else "No image
evidence provided.",
"",
"Return only the JSON object. Example:",
'{"verdict": "Supported", "justification":
"Because [TEXT_0] shows ... and [IMAGE_1]
corroborates ..."}'
```

C Fine-tuning parameters

Parameter	Value
Model (base)	`unsloth/Qwen3-VL-8B-Instruct` `unsloth/Qwen3-VL-2B-Instruct`
PEFT / LoRA method	LoRA-style PEFT
LoRA rank (r)	16
LoRA α	16
LoRA dropout	0.0
LoRA bias	none
Tokenizer	loaded with checkpoint above
Optimizer	`adamw_8bit`
Learning rate	$\mathrm{LR} = 2 \times 10^{-4}$
Weight decay	0.001
LR scheduler	linear
Warmup steps	5
Per-device train batch size	16
Gradient accumulation steps	4
Num. training epochs	1
Random seed	3407
Max sequence length	50000

Table 6: The configuration follows default values from the unsloth framework (Daniel Han and team, 2023)

After fine-tuning the LoRA adapter, the adapter was merged with the base model for further usage.

C.1 Prompt used in Fine-tuning

We used the following prompt during our fine-tuning of Qwen3-VL models, which also documents the information we included from the provided training data:

```
"You are a professional fact-checker. You will
receive an image or images and a corresponding
claim, plus a question asking you to evaluate
the claim. Your task: using the evidence
(images + metadata + claim + question), produce
a JSON object with exactly two keys:
\"justification\" and \"verdict\".\n\n"
"VERDICT LABELS (choose exactly one):\n"
" - Supported: the evidence supports the
claim.\n"
" - Refuted: the evidence contradicts the
claim.\n"
" - Not Enough Evidence: there is
insufficient evidence to either support or
refute the claim.\n"
" - Conflicting Evidence/Cherry-picking:
there exists both supporting and refuting
evidence (or evidence has been cherry-picked).
\n\n"
"IMPORTANT: Output exactly one JSON object
and nothing else.\n\n"
f"Claim: {claim_text}\n"
f"Question (current): {question_text}\n"
f"Question Type (current): {',
'.join(q_question_type) if
isinstance(q_question_type,
(list,tuple)) else q_question_type}\n"
f"Answers (for current question):
{json.dumps(q_answers, ensure_ascii=False)}
\n\n"
f"All Questions (full list):
{json.dumps(questions_summary,
ensure_ascii=False)}\n\n"
f"Fact-checking strategies:
{json.dumps(entry_fact_checking_strategies,
ensure_ascii=False)}\n"
f"Modality: {entry_modality}\n"
f"Refuting reasons: {json.dumps(
entry_refuting_reasons,
ensure_ascii=False)}\n"
f"Image misuse types: {json.dumps(
entry_image_misuse_types,
ensure_ascii=False)}\n"
f"Image used: {entry_image_used}\n\n"
f"Location: {location}\n"
f"Date: {date}\n"
f"Media source: {media_source}\n"
f"Original claim URL:
{original_claim_url}\n"
f"Reporting source: {reporting_source}\n"
```