

Multimodal Claim Extraction for Fact-Checking

Joycelyn Teo^{1,*}, Rui Cao², Zhenyun Deng², Zifeng Ding²
Michael Sejr Schlichtkrull³, Andreas Vlachos²

¹Defence Science and Technology Agency, Singapore

²University of Cambridge, UK, ³Queen Mary University of London, UK

tlimeijo@dsta.gov.sg, {rc990,zd302,zd320,av308}cam.ac.uk

m.schlichtkrull@qmul.ac.uk

Abstract

Automated Fact-Checking (AFC) relies on claim extraction as a first step, yet existing methods largely overlook the multimodal nature of today’s misinformation. Social media posts often combine short, informal text with images such as memes, screenshots, and photos, creating challenges that differ from both text-only claim extraction and well-studied multimodal tasks like image captioning or visual question answering. In this work, we present the first benchmark for multimodal claim extraction from social media, consisting of posts containing text and one or more images, annotated with gold-standard claims derived from real-world fact-checkers. We evaluate state-of-the-art multimodal LLMs (MLLMs) under a three-part evaluation framework (semantic alignment, faithfulness, and decontextualization) and find that baseline MLLMs struggle to model rhetorical intent and contextual cues. To address this, we introduce MICE, an intent-aware framework which shows improvements in intent-critical cases.¹

1 Introduction

The spread of misinformation on social media is increasingly multimodal, amplifying both its credibility and viral reach (Akhtar et al., 2023). Recent studies estimate that over one-third of debunked claims involve both text and images, highlighting the need for automated systems capable of reasoning across modalities (Zeng et al., 2024; Van Der Meer et al., 2025). This shift has motivated interest in multimodal fact-checking (Alam et al., 2025), where prior work has primarily focused on claim detection (Cheema et al., 2022) and verification (Yao et al., 2023; Braun et al., 2025). How-

*This work was partially done during Joycelyn’s research visit at Cambridge.

¹The MMCE dataset is available at <https://huggingface.co/datasets/joycelynt/MMCE>, and the code can be accessed at <https://github.com/jt9080/MMCE-Paper>.

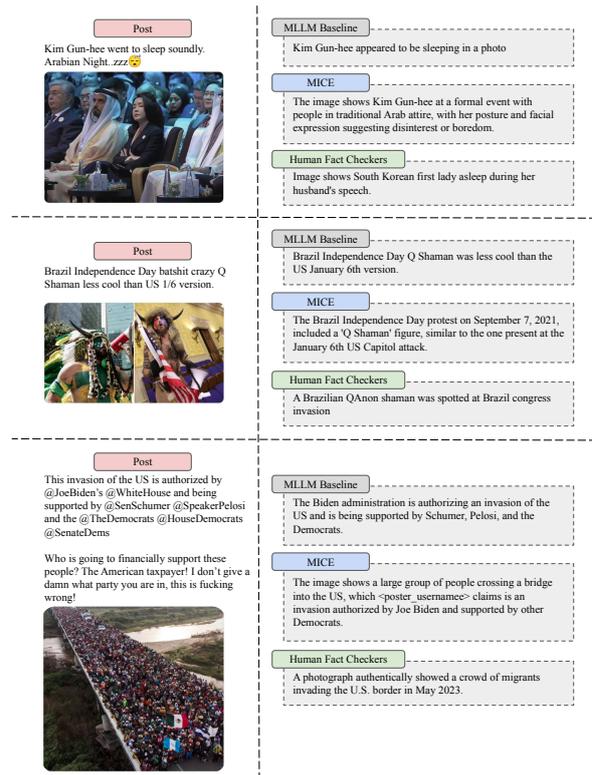


Figure 1: Examples of claim extraction from the MMCE dataset illustrating how baseline MLLMs (with in-context learning) underperform on intent-heavy multimodal posts. These cases highlight situations where structured reasoning (MICE) can help, but they are not representative of our dataset as a whole.

ever, multimodal claim extraction remains largely unexplored, despite being a crucial step in the fact-checking process (Hassan et al., 2015).

Extracting claims from multimodal content is challenging as systems must interpret informal language which often contains errors, integrate complementary or contradictory signals from images, and remove irrelevant context while preserving the factual core of a post (Wang et al., 2025). Figure 1 shows examples in which the Gemini 2.0 Flash baseline fails to capture these nuances, as it tends to extract the claim literally as the text states or the

image depicts, rather than inferring any underlying implication or context about the scenario.

To address this gap, in this paper, we make the following contributions:

1. We introduce MMCE (MultiModal Claim Extraction for fact checking), a new real-world benchmark dataset of 732 social media image-text posts paired with gold-standard claims, derived from fact-checking sources.
2. We introduce a scalable three-part evaluation framework for multimodal claim extraction, measuring semantic alignment with a manually annotated claim, faithfulness to the source, and decontextualization (i.e. whether a claim contains all necessary contextual information).
3. We provide a systematic study of MLLMs for claim extraction, showing that while they capture surface information, they often miss rhetorical intent and contextual cues.
4. We introduce the Multimodal Intent-aware Claim Extraction framework (MICE), which is a practical framework that combines vision-based semantic analysis with MLLMs to perform intent- and context-aware claim extraction. MICE breaks down the claim extraction process, extracting visual information and modeling underlying intent and context before performing claim extraction, so that implicit claims can be surfaced rather than literal descriptions. Our analysis shows that MICE can act as a mitigation strategy for MLLMs when handling high-nuance, intent-critical claims.

2 MMCE (Multimodal Claim Extraction)

Dataset We construct MMCE using post-claim pairs extracted from AVerImaTeC, a recently proposed dataset for real-world image-text claim verification based on data from fact-checking organizations (Cao et al., 2025). The dataset focuses on out-of-context (OOC) image-text claims because recent studies had observed that context-manipulated claims are the most common type of media-based misinformation (Dufour et al., 2024). Additionally, extraction for OOC claims is the most different from traditional vision-language tasks, such as image captioning, where the image and the text align. The task deviates from the pre-training objectives of MLLMs, presenting a significant challenge.

Each post-claim pair in the dataset consists of the original text and image(s) from the social media post, as well as the extracted claim from the corresponding fact-checking article. To do this, we first filter out data without links to the original social media posts, as well as data with claims that do not originate from social media sites. Next, we retrieve the source text directly from the linked social media posts. The final dataset contains 732 real-world post-claim pairs.²

For our experiment, we also curated a subset of 50 intent-critical claim examples that represent the failure modes that MICE is engineered to mitigate, to further analyze the utility of the intent-aware approach.

Evaluation In evaluating multimodal claim extraction, our goal is to go beyond simple similarity with a gold reference and instead capture multiple aspects of claim quality that matter for downstream verification in practice. Thus, we draw on the insights from previous work, which demonstrate that LLMs can serve as effective evaluators, offering scalable, consistent, and context-sensitive judgments (Liu et al., 2023; Fu et al., 2024; Muhamed, 2025; Es et al., 2024). Moreover, existing work on the evaluation of text-only claim extraction goes beyond evaluating similarity, but also considers faithfulness to and coverage of the original content, as well as whether the claim contains all necessary contextual information (Metropolitansky and Larson, 2025; Ullrich et al., 2025; Deng et al., 2024). These more comprehensive evaluation protocols provide a fuller picture of how well an extraction system supports downstream verification.

Motivated by these insights, we adopt an evaluation scheme consisting of both reference-based and reference-free metrics: (1) Reference-based evaluation uses an LLM to judge the degree of semantic alignment between the generated claim and the gold reference claim; (2) Entailment evaluation measures the extent to which the generated claim is faithful to the original social media post from which it was derived, while assuming that the post is true; (3) Decontextualization evaluation determines whether the extracted claim is interpretable as a stand-alone factual statement, without requiring additional context.

Our choice of model for evaluation is informed by previous studies, which have shown that Gemini

²The final dataset includes 618 train / 114 dev pairs from AVerImaTeC; the original test split was not publicly available.

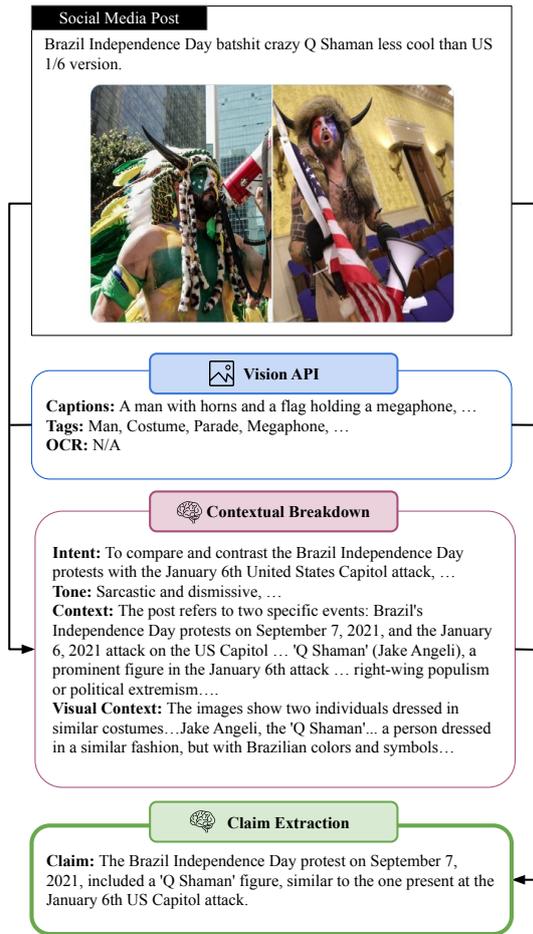


Figure 2: Overview of the MICE framework, which leverages visual understanding tools and MLLMs to reason across modalities.

models correlate relatively more with human assessments compared to other models (Akhtar et al., 2025; Gu et al., 2025). As such, we use Gemini 2.5 Flash Lite as a judge.

3 MICE (Multimodal Intent-aware Claim Extraction)

We introduce MICE, a novel approach for extracting factual claims from social media posts. While baseline MLLMs can produce fluent claim outputs, our analysis shows that they often interpret text and image content too literally without inferring underlying intent and miss contextual or rhetorical cues. To address these shortcomings, MICE first extracts visual semantics from the image of the social media post to improve visual understanding, which is an approach supported by previous work showing that combining vision APIs with LLMs is effective for multimodal reasoning (Khademi et al.,

2023). Next, MICE employs a contextual breakdown module to model intent, tone, and situational context, before integrating all signals to generate concise and accurate factual claims. The overall architecture is illustrated in Figure 2.

Vision API Given an image-text social media post, we first apply the off-the-shelf Azure Vision API to obtain structured descriptions of visual content. Specifically, we extract (1) Dense captions: natural language descriptions of salient objects, attributes, and relations, (2) OCR text: textual elements embedded in the image, and (3) Tags: high-level labels (e.g. protest, flooding).

Contextual Breakdown To capture intent and rhetorical context often present in social media posts, we prompt an MLLM for attributes which align with previous NLP-based frameworks for deception analysis, which models intention, tone, and contextual factors as core dimensions of deceptive content (Rani et al., 2025). To strike a balance between granular taxonomy and to avoid over-constraining the model outputs, we categorize the contextual breakdown into four crucial attributes: (1) Intention captures what the author tries to achieve with the post (e.g. joking, criticizing), which helps distinguish factual claims from rhetorical or performative speech, (2) Tone reflects the emotional or stylistic delivery (e.g. sarcastic, ironic, outraged), which signals whether a claim should be interpreted literally or as exaggeration, (3) Context accounts for the surrounding situational and cultural cues (e.g. ongoing events/conflicts, pop culture), which provide the background necessary to construct a sound claim, and (4) Image Context identifies specific visual elements (people, objects) shown in accompanying images, enabling the model to ground its understanding in both textual and visual information.

Claim Extraction Finally, the claim extractor MLLM receives the post text and image(s), the visual representation, and the contextual information. It is prompted to produce factual claims with In-Context Learning (ICL), in which similar demonstrations are retrieved from the training data using BM25 (Robertson and Zaragoza, 2009) under a leave-one-out strategy to prevent data leakage.

4 Experiments

Methods We evaluate a range of methods to establish performance across a diverse set of state-of-

Method	Model (used in the method)	Reference-Based (1–4) (↑)	Entailment (%)		Decontextualization (%)	
			Strict (↑)	Lenient (↑)	Strict (↑)	Lenient (↑)
MMCE dataset						
<i>LLM-based Evaluation</i>						
MLLM (text input only)	Gemini 2.0 Flash	2.80	80.0	85.9	96.9	99.7
	Qwen2.5 VL 32B Instruct	2.85	77.7	86.3	96.4	99.6
	GPT 4o Mini	2.83	80.6	88.0	97.4	99.7
MLLM	Gemini 2.0 Flash	3.11	75.5	85.4	97.9	99.9
	Qwen2.5 VL 32B Instruct	3.14	77.3	86.6	98.5	100.0
	GPT 4o Mini	3.15	74.4	85.2	97.8	99.9
MLLM with ICL	Gemini 2.0 Flash	3.21	70.5	82.6	98.4	100.0
	Qwen2.5 VL 32B Instruct	3.24	71.8	83.6	98.4	100.0
	GPT 4o Mini	3.22	69.9	82.4	98.1	100.0
MICE	Gemini 2.0 Flash	3.25	54.9	74.6	98.8	100.0
	Qwen2.5 VL 32B Instruct	3.15	52.7	78.9	98.1	100.0
	GPT 4o Mini	3.13	65.8	83.9	98.5	100.0
Intent-Critical Subset of MMCE						
<i>Human Evaluation</i>						
MLLM with ICL	Gemini 2.0 Flash	2.60	76.0	96.0	83.0	99.0
MICE	Gemini 2.0 Flash	3.31	81.0	97.0	92.0	99.0
<i>LLM-based Evaluation</i>						
MLLM with ICL	Gemini 2.0 Flash	2.08	68.0	82.0	94.0	100.0
MICE	Gemini 2.0 Flash	3.56	48.0	68.0	100.0	100.0

Table 1: Experimental results in the full MMCE dataset, as well as the intent-critical subset of it. The bold values represent the best performing method for each dataset, and for each evaluation method (human and LLM-based). Scores in the *Reference-Based* column are on a 1–4 scale (1 = lowest, 4 = highest). *Entailment* and *Decontextualization* are a 3-class categorical value, and is shown as strict (% fully entailed / fully decontextualized) and lenient (% fully or partially entailed / partially decontextualized).

the-art MLLMs: Google Gemini 2.0 Flash (DeepMind, 2024), Qwen2.5 VL 32B Instruct (Bai et al., 2025), and OpenAI GPT 4o Mini (OpenAI, 2024). We compare four methods: (1) MLLMs with text input only to evaluate the significance of visual information, (2) MLLMs with image-text inputs, (3) MLLMs with image-text inputs and ICL examples selected using BM25, and (4) the intent-aware framework, MICE.

Main results Table 1 summarizes the results across the three evaluation metrics. In reference-based scoring, we find that MLLMs with image-text inputs consistently outperform instances with text-only inputs, underscoring the importance of incorporating visual signals for claim extraction in noisy social media posts. The further gains from MLLM with in-context learning (ICL) suggest that multimodal intent is best captured when models are guided by examples. In the subset of intent-critical cases, both human and LLM-based evaluations agree that the MICE framework improves reference-based scores, though its utility is not reflected across the entire dataset because of other challenges. E.g, when interpreting an image about the packaging of COVID-related medication, MLLMs and MICE fail to spot the insinuation that COVID is the same as influenza, which is indica-

tive of the challenges MMCE poses to MLLMs more broadly (see error analysis in Appendix H).

MICE also improves decontextualization scores for Gemini 2.0 flash and GPT 4o Mini, producing claims that are more stand-alone and interpretable without the original post. This is crucial for downstream fact-checking pipelines, where claims are often checked in isolation. Our experiments also reveal that improvements in reference alignment and decontextualization introduce a trade-off with entailment. This is likely because explicitly modeling intent and context encourages the model to abstract away from the literal text and image.

Human Alignment with Automated Evaluations

To validate the LLM-based evaluation metrics and verify the utility of the MICE approach in handling intent-critical claims, we had 4 expert annotators independently score the claims extracted from the intent-critical subset (details in Appendix G), and calculated the agreement between the human evaluators and the LLM scorer. The agreement is measured using Krippendorff’s α (Krippendorff, 2013) and Spearman’s ρ (Spearman, 2010). Results (Appendix B1) show that reference-based and decontextualization alignment achieved moderate correlation with humans ($\alpha = 0.59$ and $\alpha = 0.54$ respectively), but entailment score achieved a lower

correlation ($\alpha = 0.07$). This suggests that judging entailment involves a deeper inferential reasoning of the post compared to the other metrics, leading to a larger divergence from human judgment.

Experiment on Temporal Leakage To further examine the extent to which MLLMs rely on pre-training knowledge when generating claims, we compared Gemini 2.0 flash (training cutoff: June 2024) (DeepMind, 2024) with Gemini 2.5 flash (training cutoff: January 2025) (DeepMind, 2025). Experiments on 50 randomly sampled image–text claims posted between July and December 2024 (Appendix C) showed that Gemini 2.5 flash did not have a significant improvement in scoring metrics (Table C2). On closer analysis of the improved claims, we also found no clear evidence that Gemini 2.5 flash relied on additional pre-training knowledge unavailable to Gemini 2.0 flash.

5 Conclusion

In this work, we introduce MMCE, the first dataset for multimodal claim extraction³ from social media, and established a three-part evaluation framework that measures semantic alignment, faithfulness, and decontextualization. Our analysis demonstrates that while baseline MLLMs can extract literal claims, they often miss the rhetorical intent and contextual framing crucial for understanding social media content. To address this gap, we propose MICE, an intent-aware framework that improves performance in these nuanced, intent-critical cases. Overall, this research contributes a new benchmark and an intent-aware framework, supporting the ongoing development of more effective automated tools for fact-checking.

Acknowledgments

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

³During the reviewing process we became aware of concurrent work by (Geng et al., 2026) that also proposes a dataset for this task.

Limitations

We acknowledge that claim extraction can inherently be a subjective task, whereby extracted claims with different semantic meaning can still be considered valid and check-worthy. Our rationale is to model the claim extraction process as closely as possible to that of professional fact checkers, which motivated the decision to curate a dataset from real-world fact-checking articles.

Moreover, the dataset we curated explicitly focuses on out-of-context images, where we deemed it the most urgent to focus our efforts on these types of claim. However, this excludes image-text claim types that could also benefit from the MICE framework, such as memes. In future work, we propose to extend the experiments to other types of image-text claims for diversity.

Lastly, due to resource constraints, we performed human evaluation on 100 generated claims. Although this sample size provided a representative assessment of the claims’ quality, it limits a more comprehensive statistical analysis of the framework’s performance variability across different claim domains and styles. A larger scale human evaluation would be beneficial for future iterations of this work to improve statistical reliability and reproducibility.

Ethical Considerations

We rely on fact-checks from real-world fact-checkers to develop and evaluate our models. Nevertheless, as any dataset, it is possible that it contains biases which influenced the development of our approach. Given the societal importance of fact-checking, we advise that any automated system is employed with human oversight to ensure that the fact-checkers fact-check appropriate claims.

References

- Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. [Multimodal automated fact-checking: A survey](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.
- Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2025. [Ev2r: Evaluating evidence retrieval in automated fact-checking](#).
- Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna

- Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venkatesh V. 2025. [The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and & retrieval](#). In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, page 467–478, Berlin, Heidelberg. Springer-Verlag.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. [Qwen2.5-vl technical report](#).
- Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. [DEFAME: Dynamic evidence-based FAct-checking with multimodal experts](#). In *Forty-second International Conference on Machine Learning*.
- Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. [Averimatec: A dataset for automatic verification of image-text claims with evidence from the web](#).
- 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.
- DeepMind. 2024. [Gemini 2.0 flash model card](https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf). <https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf>.
- DeepMind. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#).
- Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. [Document-level claim extraction and de-contextualisation for fact-checking](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.
- Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Duffield, 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](#).
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. [GPTScore: Evaluate as you desire](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2026. [M4fc: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset](#).
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#).
- Naemul Hassan, Chengkai Li, and Mark Tremayne. 2015. [Detecting check-worthy factual claims in presidential debates](#). In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, page 1835–1838, New York, NY, USA. Association for Computing Machinery.
- Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. [MM-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, Singapore. Association for Computational Linguistics.
- K. Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Dasha Metropolitansky and Jonathan Larson. 2025. [Towards effective extraction and evaluation of factual claims](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6996–7045, Vienna, Austria. Association for Computational Linguistics.
- Aashiq Muhamed. 2025. [Ccrs: A zero-shot llm-as-a-judge framework for comprehensive rag evaluation](#).
- OpenAI. 2024. [Gpt-4o system card](#).
- Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2025. [SEPSIS: I can catch your lies – a new paradigm for deception detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for*

Computational Linguistics (Volume 4: Student Research Workshop), pages 97–128, Vienna, Austria. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.

C Spearman. 2010. [The proof and measurement of association between two things](#). *International Journal of Epidemiology*, 39(5):1137–1150.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2025. [Claim extraction for fact-checking: Data, models, and automated metrics](#).

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](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31274–31291, Vienna, Austria. Association for Computational Linguistics.

Haoran Wang, Aman Rangapur, Xiong Xiao 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.

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, SIGIR '23*, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. 2024. [Multimodal misinformation detection by learning from synthetic data with multimodal LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10467–10484, Miami, Florida, USA. Association for Computational Linguistics.

A MMCE Dataset

The breakdown of the social media sites of the data from MMCE is shown in table A1.

Social Media Site	Number of Claims
X (Formally Twitter)	360
Facebook	319
Instagram	32
Reddit	13
Telegram	2
Weibo	2
Band (Naver)	1
Flickr	1
Truth Social	1

Table A1: Breakdown of social media sites

Although we converted the original post text to English for our experiments, we also retained the original language of the social media posts for future experiments. Table A2 shows the breakdown of the language of the original posts in MMCE.

Original Post Language	Number of Claims
English	551
Hindi	100
Korean	14
Urdu	10
Chinese	8
Bengali	7
Thai	6
Sinhala	6
Khmer	7
Burmese	6
Filipino	4
Amharic	4
Tamil	3
Indonesian	2
Punjabi	2
Telugu	2
Kannada	1
French	1
Pashto	1
Swahili	1

Table A2: Breakdown of the language of the original posts in MMCE. Some posts use more than one language. In those cases, we count it under all the languages used.

B Human Alignment Analysis

Table B1 reports detailed agreement statistics between LLM-based evaluations and human annotators, as well as human–human agreement.

Due to resource constraints, one annotator labeled all 100 examples; the other three split the set (approximately 33–34 claims each).

C Temporal Leakage Analysis

Data for the temporal leakage analysis was drawn from four fact-checking sources via the Google Fact Check API. The breakdown of the source URLs of the data used in the temporal leakage analysis is shown in table C1.

Fact Checking Site	Number of Claims
factcheck.afp.com	30
leadstories.com	13
politifact.com	5
fullfact.org	2

Table C1: Breakdown of fact checking sites used in the temporal leakage experiments

The results of the temporal leakage experiment are shown in table C2.

D Model Implementation and Resources

For few-shot settings, we set the number of shots to be 5 to strike a balance between providing sufficient demonstrations, and to prevent exceeding the context length limit.

For experiments with Qwen2.5 VL 32B Instruct, we faced a 30-image limit per input. In the handful of cases where the limit is exceeded (due to the images from the few shot demonstrations), we limit the number of few-shot inputs to fit the image constraints.

All experiments were conducted using API access to commercial and open-source multimodal LLMs via the OpenRouter⁴ platform. We report results for Gemini 2.0 Flash, Gemini 2.5 Flash, Qwen2.5 VL 32B Instruct, and GPT-4o Mini. Since these models are hosted services, we do not control the underlying hardware; however, we record the model versions and set temperature to zero to facilitate reproducibility.

⁴<https://openrouter.ai/>

Method	Model	Reference-Based	Entailment	Decontextualization
LLM-Human	Krippendorff’s α	0.59	0.07	0.54
	Spearman ρ	0.56	0.11	0.41
	Agreement (%)	46.5	50.0	89.5
Human-Human	Krippendorff’s α	0.67	0.14	0.80
	Spearman ρ	0.70	0.30	0.81
	Agreement (%)	55.0	69.0	94.0

Table B1: Human–LLM and human–human alignment statistics across the three evaluation metrics used in the experiment.

Model	Reference-Based (1–4) (↑)	Entailment (%)		Decontextualization (%)	
		Strict (↑)	Lenient (↑)	Strict (↑)	Lenient (↑)
Gemini 2.0 Flash	3.08	88.0	90.0	94.0	98.0
Gemini 2.5 Flash	3.10	86.0	86.0	84.0	90.0

Table C2: Results of the temporal leakage experiment, on 50 post-claim pairs that surfaced after the training cut-off for Gemini 2.0 Flash, but before the training cut-off for Gemini 2.5 Flash. Scores in the *Reference-Based* column are on a 1–4 scale (1 = lowest, 4 = highest). *Entailment* and *Decontextualization* are shown as strict (% fully entailed / fully decontextualized) and lenient (% fully or partially entailed / partially decontextualized).

E Licensing Information

If accepted, we will publicly release and maintain the dataset and baseline code, which will be licensed under the CC BY-NC 4.0 license.

F Prompts

Here we provide the prompts used in our experiments. This includes the prompts used in the LLM-based evaluation, as well as the prompts used in the baseline approaches and the MICE framework.

For the MICE framework, the prompt used in the claim extraction stage is a concatenation of the image-text input prompt (F.5), the output of the Vision API, and the output of the contextual extraction (F.6).

F.1 Prompt for reference-based evaluation

```
# MISSION
You will get two claims below, a
generated claim and a reference claim.
Your task is to perform a comprehensive
similarity assessment between the
generated claim and the reference
claim, in the context of fact checking.
Provide a similarity score from 1 to 4.

# CRITICAL INSTRUCTIONS:
- Be CONSISTENT in your scoring.
Similar claims should always receive
similar scores.
- Focus on the core implied
meaning/content of the sentence,
wording differences should be
acceptable.
```

- Extra details that don’t contradict the core claim is beneficial. Additional specific details (e.g. names, dates, locations, numbers) on the core factual assertion is a good thing. Do not penalize for verbosity and specificity.

```
# Scoring Guidelines:
- 1: Completely different, no overlap
in themes, topics, or entities
mentioned.
- 2: Minimal similarity in themes,
topics, or entities mentioned,
sentences have different meanings. The
core factual statement to be fact
checked is different.
- 3: Partial alignment in message
conveyed, with significant differences
that could potentially affect
downstream fact checking.
- 4: Strong conceptual similarity with
minor variations or near-identical
meaning.
```

```
# RESPONSE FORMAT
Return the response in the following
JSON format:
““json
{
  "score": 3,
  "reasoning": "Brief explanation of
why this score was given"
}
““

# INPUT
Generated Claim: <generated claim>
Reference Claim: <reference claim>
```

F.2 Prompt to evaluate entailment

```
# MISSION
You will be given a social media post
via the text and image of the post, we
well as a candidate claim extracted
from the post. Your task is to assess
whether the claim is fully faithful to
and entailed by the combined content of
the image and text. This means that
assuming the social media post is true,
the extracted claim must also be true.
```

Do NOT check if the claim is true in reality, only whether it is faithful to the content of the image and text of the post.

Ignore whether the correct factual content had been extracted, focus on whether the extracted sentence is faithful (i.e. no hallucinations).

```
# EVALUATION CRITERIA
Classify the claim into one of three
categories:
- **entailed** : The claim is fully
aligned with the post content without
any contradictions, hallucinations or
unsupported additions.
- **partially_entailed** : The claim is
partially aligned with the post content
but contains minor variations or
additional context not stated or
implied in the post.
- **not_entailed** : The claim contains
significant misaligned inferences,
exaggerations beyond what's stated,
major contradictions, hallucinations,
or completely misaligns with the post
content.
```

```
# INPUT
Generated Claim: <generated claim>
Text: <social media post text>
Image(s): <social media post image(s)>
```

F.3 Prompt to evaluate decontextualization

```
# MISSION
You will be given a candidate claim
that was extracted from a social media
post. Your task is to assess whether
this claim is understandable in
isolation, without access to the
original post or any external context.
A decontextualized claim should be
fully interpretable and self-contained
to an average reader, who has no
knowledge of the post.
```

```
# CRITICAL INSTRUCTIONS:
- Focus only on clarity and
completeness of meaning. Do not check
whether the claim is factually true or
faithful to the post, only whether the
claim can stand alone and be understood
independently.
```

```
# Scoring Guidelines:
- **fully_decontextualized** :
Understandable in isolation. The claim
is completely self-contained,
unambiguous, and requires no edits to
be understood on its own. (Example: The
mayor of NYC announced a new recycling
program on June 1, 2024.)
- **partially_decontextualized** : The
claim is mostly clear and contains some
context, but has gaps, vague references
or unresolved pronouns. The claim could
benefit from some edits. (Example:
Vaccination rates rose after that. ->
could be rewritten to -> Vaccination in
the UK rates rose after the 2023
campaign.)
- **not_decontextualized** : Not
understandable in isolation. The claim
cannot be interpreted on its own; key
entities, referents, or context are
missing. Major rewriting is needed.
(Example: He did something yesterday.)
```

```
# INPUT
Generated Claim: <generated claim>
```

F.4 Prompt for text-only input for claim extraction

```
# MISSION
You are an expert fact-checking analyst
specializing in social media content
verification. Your primary objective is
to precisely extract and articulate the
core factual claim(s) from the given
text.
```

```
# CONTEXT ANALYSIS
Before extracting the claim, perform a
comprehensive context analysis:
- Examine the full text carefully
- Consider the platform type and its
typical communication style
- Identify potential implicit or
explicit claims
```

```
# CLAIM EXTRACTION METHODOLOGY
1. Identify Potential Claims:
- Look for definitive statements
- Detect implied assertions
- Recognize potentially misleading
or exaggerated claims
2. Claim Criteria:
- Clarity: Can the claim stand alone
and be understood without the
original context?
- Specificity: Does the claim
capture the most significant factual
assertion?
- Verifiability: Does the claim
provide enough detail to enable
fact-checking?
3. Claim Refinement Process:
- Remove subjective language
- Distill the core factual assertion
```

- Ensure the claim is neutral and objective

CLAIM SELECTION STRATEGY

- Always try to extract just one main claim first
- If the text contains one main factual assertion, extract only that claim
- If multiple statements can be combined into one coherent claim, do so

Multiple claims should only be used when:

- The text contains completely separate factual statements about different topics that cannot be combined
- Each claim is independently verifiable and fact-checkable
- Combining them would create a confusing or overly complex claim

ADDITIONAL CONSIDERATIONS

- If multiple potential claims exist, first try to identify the most significant or impactful one
- If the claim is ambiguous, provide the most reasonable interpretation based on context
- Avoid introducing personal bias or speculation
- Always prioritize extracting a single, comprehensive claim over multiple separate claims

RESPONSE FORMAT

Return the response in the following JSON format:

```

{
  "claims": ["main claim"]
}

```

INPUT

Extract the claim(s) from the following text: <social media post text>

F.5 Prompt for image-text input for claim extraction

MISSION

You are an expert fact-checking analyst specializing in social media content verification. Your primary objective is to precisely extract and articulate the core factual claim(s) from the given text and accompanying image.

CONTEXT ANALYSIS

Before extracting the claim, perform a comprehensive context analysis:

- Examine the full input text and image carefully. Consider how the image contributes to the messaging
- Consider the platform type and its typical communication style
- Identify potential implicit or explicit claims

CLAIM EXTRACTION METHODOLOGY

1. Identify Potential Claims:

- Look for definitive statements
- Detect implied assertions
- Recognize potentially misleading or exaggerated claims

2. Claim Criteria:

- Clarity: Can the claim stand alone and be understood without the original context?
- Specificity: Does the claim capture the most significant factual assertion?
- Verifiability: Does the claim provide enough detail to enable fact-checking?

3. Claim Refinement Process:

- Remove subjective language
- Distill the core factual assertion
- Ensure the claim is neutral and objective
- Consider whether the image alters, reinforces, or creates the perceived claim

CLAIM SELECTION STRATEGY

- Always try to extract just one main claim first
- If the text contains one main factual assertion, extract only that claim
- If multiple statements can be combined into one coherent claim, do so

Multiple claims should only be used when:

- The text contains completely separate factual statements about different topics that cannot be combined
- Each claim is independently verifiable and fact-checkable
- Combining them would create a confusing or overly complex claim
- The image introduces additional factual assertions that cannot be combined with the text claims

ADDITIONAL CONSIDERATIONS

- If multiple potential claims exist, first try to identify the most significant or impactful one
- If the claim is ambiguous, provide the most reasonable interpretation based on context
- Avoid introducing personal bias or speculation
- Always prioritize extracting a single, comprehensive claim over multiple separate claims
- Consider whether the image alters, reinforces, or creates the perceived claim

RESPONSE FORMAT

Return the response in the following JSON format:

```

{
  "claims": ["main claim"]
}

```

```

}
'''

# INPUT
Extract the claim(s) from the following
text: <social media post text>

```

F.6 Prompt to extract contextual insights

```

# MISSION
Analyze this social media post and
provide contextual insights to help
identify the main factual claim.

# CONTEXTUAL ANALYSIS
Focus on these key insights:
1. INTENT: What's the main purpose of
the post?
(inform/persuade/entertain/satire/etc.)
2. TONE: What's the emotional tone of
the post?
(serious/humorous/sarcastic/anger/etc.)
3. CONTEXT: What real-world
events/issues does this relate to?
Include specific details like dates,
locations, people, organizations.
4. VISUAL_CONTEXT: What specific
people, objects, locations, or events
are shown in the image that provide
context for the claim?

# RESPONSE FORMAT
Return your analysis as a JSON object
with the following structure:
'''json
{
  "intent": "description of the
poster's main purpose",
  "tone": "description of the
emotional tone",
  "context": "brief context with
specific details about real-world
events/issues",
  "visual_context": "description of
what's shown in the image that
provides context"
}
'''

```

G Annotation Guidelines

The annotators involved in the human evaluation are also co-authors but were unaware of which system generated which claim. Here, we provide the annotation guidelines provided to those annotators, alongside an excel sheet containing the intent-critical subset of the 100 post-claim pairs, as well as the images relevant to the posts.

G.1 Instructions given to the annotators

There are 100 generated claims in the excel file, and for each claim, the original social media post (text+image(s)) and the reference claim, are provided. Fill in the green

columns (E and F): reference-based score and entailment score.

Here are the guidelines for the scoring:

Reference-based
[Look at columns C and D only]
Given 2 claims, a generated claim and a reference claim, perform a comprehensive similarity assessment between the generated claim and the reference claim. Provide a similarity score from 1 to 4.

Focus on the core meaning and factual content, not minor wording differences. Ignore extra details that don't change the core factual claim, no need to penalize for that.

Scoring Guideline (1-4 scale)
1: Completely different, no overlap in themes, topics, or entities mentioned.
2: Minimal similarity in themes, topics, or entities mentioned, sentences have different meanings. The core factual statement to be fact checked is different.
3: Partial alignment in message conveyed, with significant differences that could potentially affect downstream fact checking.
4: Strong conceptual similarity with minor variations or near-identical meaning.

Entailment
[Look at columns A, B and D only]
Given a social media post (text and image(s)), as well as a claim extracted from the post, assess whether the claim is fully faithful to and entailed by the combined content of the image and text. This means that assuming the social media post is true, the extracted claim must also be true.

Ignore whether the correct factual content had been extracted, focus on whether the extracted sentence is faithful to the post (i.e. no hallucinations).

Scoring Guideline (3-class classification)
1. entailed: The claim is fully aligned with the post content without any contradictions, hallucinations or unsupported additions.
2. partially_entailed: The claim is partially aligned with the post content but contains minor variations or additional context not stated or implied in the post.
3. not_entailed: The claim contains significant misaligned inferences, exaggerations beyond what's stated, major contradictions, hallucinations, or completely misaligns with the post content.

Decontextualization

[Look at column D only]

You will be given a candidate claim that was extracted from a social media post. Your task is to assess whether this claim is understandable in isolation, without access to the original post or any external context. A decontextualized claim should be fully interpretable and self-contained to an average reader, who has no knowledge of the post.

Scoring Guideline (3-class classification)

1. fully_decontextualized:

Understandable in isolation. The claim is completely self-contained, unambiguous, and requires no edits to be understood on its own. (Example: The mayor of NYC announced a new recycling program on June 1, 2024.)

2. partially_decontextualized: The claim is mostly clear and contains some context, but has gaps, vague references or unresolved pronouns. The claim could benefit from some edits. (Example: Vaccination rates rose after that. -> could be rewritten to -> Vaccination in the UK rates rose after the 2023 campaign.)

3. not_decontextualized: Not understandable in isolation. The claim cannot be interpreted on its own; key entities, referents, or context are missing. Major rewriting is needed. (Example: He did something yesterday.)

H Error Analysis

To surface the key challenges faced in image-text social media claim extraction, we select erroneous samples (scored 1 or 2 for reference-based scoring) for analysis, shown in Table H1. We also select instances where the baseline MLLM with ICL outperforms MICE in reference-based scoring, shown in Table H3, in order to highlight some potential drawbacks of MICE. Note that in cases where the system outputs more than one claim, we select the highest scoring claim based on the reference-based evaluation.

Failure Cases for both baseline MLLMs and MICE

Social Media Post	Claims	Analysis
<p>India's clothing minister - Prime Minister of Italy. Very cultured.</p> 	<p>Reference: Italy's Prime Minister Giorgia Meloni refuses to shake hands with PM Narendra Modi.</p> <p>MLLM+ICL: Image shows Narendra Modi and the Prime Minister of Italy. [2]</p> <p>MICE: The image shows Narendra Modi opening a car door for Italian Prime Minister Giorgia Meloni. [1]</p>	<p>The LLM-based approaches fail to capture the nuance of the gestures and context in the image. The phrase "Very cultured" adds a subjective commentary on behavior rather than merely describing the scene. Though MICE attempts to interpret the actions of the people in the image, it overlooks intended meaning of the post. This reflects a broader limitation in MLLMs' ability to integrate non-literal cues from language with the contextual visual semantics of human gestures.</p>
<p>Krabi people are enlightened!!! Move Forward Party is no longer a party that will overthrow the monarchy. #MoveForward #themalaengtad</p>  <p>The text overlay in the image is: "Krabi people's eyes are open! If you don't take the party forward, the party will fall."</p>	<p>Reference: The Move Forward Party's political rally in Southern Krabi province, Thailand had a very small crowd attendance.</p> <p>MLLM+ICL: The Move Forward Party is no longer a party that will overthrow the monarchy. [1]</p> <p>MICE: On April 23, 2023, the Move Forward Party held an event in Krabi, and the party is no longer viewed as wanting to overthrow the monarchy. [2]</p>	<p>The LLM-based models misinterpret the underlying sarcasm and political sentiment embedded in the post. The Thai text, combined with the image of a sparse crowd, is meant to mock the Move Forward Party's rally attendance. However, both MICE and MLLM+ICL treat the post as a straightforward political statement, failing to synthesize the ironic signals between text and visuals to understand the true rhetorical intent.</p>
<p>Packaging update 🤔</p> 	<p>Reference: COVID19 is the same as influenza.</p> <p>MLLM+ICL: Fanttest sells a COVID-19/Influenza A&B Antigen Test Kit. [1]</p> <p>MICE: The packaging for the Fanttest COVID-19/Influenza A&B Antigen Test Kit has been updated. [1]</p>	<p>The LLM-based approaches fail to identify the implicit misinformation within the post. Both MICE and MLLM+ICL interpret the post literally without recognizing the underlying insinuation that conflates two distinct diseases. The emoji also serves as a rhetorical device signaling skepticism and insinuation, which was not picked up by the systems. Consequently, their outputs remain factually descriptive but semantically shallow, demonstrating a persistent limitation in detecting implicit misinformation.</p>

Table H1: Analysis on instances where both baseline MLLMs and the MICE framework yields poor results, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score (Part 1).

Failure Cases for both baseline MLLMs and MICE

Social Media Post

The name of this Chinese firm is PMC Projects and its owner is 'Chang Chien-ting', better known as Morris Chang. Surprisingly, Morris Chang is the son of Chinese citizen Chang Chung-ling, who has been a director of Adani's companies and is a business partner of Adani's brother Vinod Adani.

2. Particulars of the holder of the significant beneficial interest:

Name of the Significant Beneficial Owner (Given name and last Name)	Chang Chien-Ting
Address and Email Id	11F-1, No.290, Xianfu Road, Taoyuan Dist., Taoyuan city, Taiwan. pmcinfra.maust@gmail.com
Date of Birth/Age	14 November 1976 / 41
Father's/ Mother's/ Spouse's name	Mr. Chang Chung Ling
Occupation	Business
Nationality	Republic of China
Passport No. (in case of foreign national)	316345230



This image is a fraud, created via AI. Supporters of Russia, you go right ahead and believe this fraud. I have no objections at all. Supporters of Ukraine, Ukraine has taken at least five villages in the last day. The Russian army in the last two days has lost immense amounts of men, far more than the daily average for this war. In the last village taken alone, fifty occupiers were slain, and four captured. Russian supporters, you won't believe me, and I absolutely do not need you to. What you believe or don't believe is immaterial. As we freed my family in Kherson, as we freed Kharkiv, so also now we free more villages from the horrible dystopia of the 'multi-polar paradise'. But you please do keep making your AI images and living in Putin's house of mirrors and illusions.



Claims

Reference: Chang Chien-Ting, an associate of the Adani Group, is a Chinese national.

MLLM+ICL: Morris Chang, owner of PMC Projects, is the son of Chang Chung-ling, who is a director of Adani's companies and a business partner of Adani's brother, Vinod Adani. [1]

MICE: Morris Chang, owner of PMC Projects, is the son of Chang Chung-ling, who is a director of Adani's companies and a business partner of Vinod Adani. [1]

Analysis

The LLM-based approaches fail to detect the deceptive framing and the nationality misattribution at the core of this post. This is likely because the post relies on background knowledge of India–China relations and domestic narratives around economic nationalism. The post falsely claims that the Adani Group's (An Indian Multinational Conglomerate) associate is linked to a "Chinese" businessman, exploiting ongoing public suspicion toward Chinese influence in Indian infrastructure projects. Both MICE and MLLM+ICL extract the literal factual components (names, companies, and family relations) but fail to recognize the misleading intent: the deliberate conflation of Taiwan ("Republic of China") with mainland China ("People's Republic of China") to frame the relationship as evidence of Chinese involvement. This illustrates a broader limitation of automated claim extraction, which in some cases would need context from real-world and real-time socio-political context.

Reference: Image Showing Ukrainian Losses Is Generated By Artificial Intelligence.

MLLM+ICL: Ukraine has taken at least five villages in the last day, and in the last village taken alone, fifty Russian soldiers were slain, and four captured. [1]

MICE: Ukraine has taken at least five villages in the last day, and the Russian army has suffered immense losses in the last two days. [1]

Both models extract a secondary claim from the lengthy post text but miss the primary, overarching claim that "This image is a fraud, created via AI" (the image is verified as legitimate). The models appear to latch onto the more concrete, detailed statements about recent military events later in the paragraph. This limitation likely arises from the claim extraction prompt, which directs the system to identify minimal or singular claims, a strategy that proves inadequate for posts containing longer, more complex paragraphs.

Table H2: Analysis on instances where both baseline MLLMs and the MICE framework yields poor results, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score (Part 2, continued from H1)).

Cases where baseline MLLMs outperforms MICE

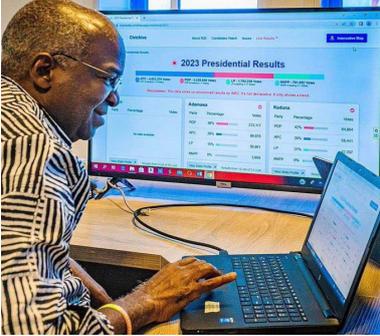
Social Media Post	Claims	Analysis
<p>Tunisian Club African CEO being interviewed by a journalist about their game with Yanga SC</p> 	<p>Reference: Image shows the CEO of Club Africain of Tunisia being interviewed after their match with Tanzania's Yanga SC.</p> <p>MLLM+ICL: The man in the photo is the CEO of Tunisian Club African being interviewed about their game with Yanga SC. [4]</p> <p>MICE: The CEO of Tunisian Club African is depicted as a man wearing a turban and traditional clothing, sitting at a desk with a laptop and an AK-47 assault rifle. [1]</p>	<p>This instance illustrates the potential drawback of applying MICE's complex reasoning to straightforward, literal posts. The MICE framework fails because of its over-emphasis on visual analysis where none is needed. The post's claim is fully contained in the text, but MICE's methodology compels it to integrate visual features. This shows that for simple, text-dominant claims, MICE's intent- and context-aware modules can introduce noise and errors rather than adding value.</p>
<p>Someone please ask Fashola how he got access to this server that has the election result data. INEC and APC rigged the election even before it started. They should come out and defend this.</p> 	<p>Reference: Nigerian housing minister Babatunde Fashola illegally accessed election data on a server prior to the official release of results.</p> <p>MLLM+ICL: A photo shows Babatunde Fashola accessing a server containing election result data. [4]</p> <p>MICE: The image shows election result data from Civichive for the 2023 Nigerian presidential election. [1]</p>	<p>MICE mentioned "result data from Civichive," a detail not mentioned in the text and likely derived from misinterpreting the website screenshot in the image. This shows that the Vision API or contextual breakdown module likely misinterpreted the UI elements in the image, whereas the baseline's more direct approach remained more faithful to the source text.</p> <p>However, the MLLM+ICL baseline claim also does not capture the key detail that Fashola allegedly illegally accessed the server. The high score is arguably inflated, as the claim omits the critical detail from the reference that the access was "illegally" obtained. The automated judge rewards the partial semantic overlap, highlighting a weakness in the reference-based scoring metric itself.</p>
<p>U.S. District Judge Aileen Cannon (on the left side in photo) has been initially assigned to oversee the case...could preside over the trial as well. Cannon, appointed by Trump in 2019, made decisions that favored him in legal skirmishes during the documents investigation last year. Her rulings were overturned on appeal. Cannon would determine, among other things, when a trial would take place and what Trump's sentence would be if he were found guilty. 😊</p> 	<p>Reference: A photo shows Aileen Cannon, the US district court judge assigned to the criminal case against Donald Trump for retaining classified documents, wearing Trump apparel outside one of the former president's rallies.</p> <p>MLLM+ICL: Judge Aileen Cannon, who was appointed by Trump in 2019 and initially assigned to oversee his case, is shown in the photo wearing pro-Trump attire. [4]</p> <p>MICE: Aileen Cannon, a Trump appointee, has been assigned to oversee the classified documents case against Donald Trump, and she previously made decisions favorable to him that were later overturned. [2]</p>	<p>Here, the MLLM+ICL baseline outperforms MICE by successfully integrating key details from both the text and the image. The reference claim's core is that Judge Cannon is wearing "Trump apparel," a visual fact explicitly mentioned in the baseline's output.</p> <p>In contrast, the MICE framework's treated the image as merely illustrative. Its reasoning process failed to grasp that the visual information was the most salient part of the claim.</p>

Table H3: Analysis on instances where the baseline MLLMs outperforms the MICE framework, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score.