Deepfakes and Their Analysis Using Large Language Models A Critical Review of Detection, Explanation and Governance

Author A. Siddharth Singh,
B. Siddharth Pratap Somvanshi
of Computer Application.
Banarasi Das University, Lucknow.
siddharthsingh161295@gmail.com
sidsomvanshi007@gmail.com

Corresponding Author, Neelam Batham Author
(Assistant Professor) School
School of Computer Applications Babu
Babu Banarasi Das University, Lucknow
Neelam.mca01@bbdu.ac.in

Abstract

Deepfakes are manipulated or created images, videos, audio and text which appear to be genuine to people. They are relatively easy to create and much more difficult to identify as fakes. The conventional method to detect deepfakes was to develop a classifier to differentiate between real and fake media. This was effective until the fakes evolved. A detector that performs well on its training data can perform poorly when confronted with a new generator, compression pipeline or language. Recently, large language models (LLMs) and multimodal LLMs (MLLMs) have entered the scene. Not only do they offer the ability to detect deepfakes, but they also provide an understanding of meaning and the capability to explain why a decision was made. This review paper examines, solely based on existing literature, how large language models are employed for deepfake content analysis. Approximately twenty-five studies were analyzed, ranging from surveys and benchmarks to zero-shot tests, a fraud case, and the 2026 amendments to India's IT Rules (one of which pertains to synthetic content). The study identified five potential applications for an LLM: zero-shot judge, explainer, fine-tuned forensic assistant, text-deepfake detector, and audio-language detector.

While general-purpose models continue to be valuable, they cannot be relied upon exclusively; language-based training enhances both comprehension and explanation; and multilingual, noisy settings continue to present considerable challenges. The paper concludes by proposing a layered detection process that utilizes LLMs in conjunction with specialized detectors, verification checks, and human reviewers.

Keywords: deepfake; large language model; multimodal; LLM forensics; explainable detection; fake voice; synthetic text; AI regulation 1.

Introduction

1.1 Background

"Deepfake" is a combination of "deep learning" and "fake." The term gained public attention around 2017, when creators began employing neural networks to exchange faces in videos and publish them online. Since then, the scope of deepfakes has expanded significantly beyond face-swapped video. All that is required to clone a voice from a brief audio recording, place an individual into a video they were never present in, or generate a realistic image of someone who does not exist is a computer and some free software. As humans are predisposed to believe what they witness and hear, such media can have detrimental effects: it can harm reputations, influence opinions, and even manipulate markets. A

widely reported incident occurring around 2024 illustrates the risks involved. An employee in finance at Arup's Hong Kong office participated in a video conference call where the purported Chief Financial Officer and several other co-workers were actually fake identities. Feeling confident, the employee transferred fifteen times approximately HK$200 million (roughly US$25 million) [26]. The fraud was discovered only when the employee contacted the Head Office at a later date. Meanwhile, Large Language Models (LLMs) are progressively used as general-purpose reasoning systems; for example, GPT-4V combines text with images and responds to open-ended inquiries across numerous tasks [9].

Therefore, it is reasonable to question whether a model capable of describing a picture and supporting its assessment regarding it can also determine if the picture is fake.

1.2 Problem Statement and Research Questions

Despite producing numerous high-scoring classifiers, the detection literature demonstrates that these classifiers become fragile when new generation methods emerge, they rarely explain their judgments, and they fail to comply with the evidentiary requirements of investigators and courts [5]. Evidence concerning LLM-assisted analysis, however, is distributed across vision, speech, language processing, and law. This paper summarizes this evidence into four research questions:

RQ1. To what extent can general-purpose or multimodal LLMs detect deepfakes without task-specific training?

RQ2. Does fine-tuning with language supervision enhance detection accuracy, localization, and explanation?

RQ3. Do the results differ based on image, video, audio, and text modalities?

RQ4. What are the most relevant risks and regulatory factors for real-world deployment?

1.3 Scope

The study considered only peer-reviewed journal articles, academic conference papers, preprints, and official or legal commentaries published through October 2026. "Analysis using LLMs" was defined broadly to encompass the detection or explanation of fake media by LLMs, as well as the detection of LLM-generated text, which was counted as a form of deepfake (although some readers may disagree).

2. Understanding Deepfakes

Early face-swap software utilized autoencoders: one network compressed a face into a coded representation, while a decoder trained on a particular individual's images reconstructed that person's face. Generative Adversarial Networks (GANs) improved visual quality by training a generator against a discriminator, and diffusion models, currently the leading technology in many consumer products, further refined generated images [6]
(Fig. 1). Scholars typically categorize facial manipulations into face swapping, re-enactment , talking-face generation, and attribute editing [6]. Face Forensics++ combined four methods (Deepfakes, Face2Face, Face Swap, and Neural Textures) to produce fake variations of the same video [1]. In terms of audio, text-to-speech and voice-conversion technologies can replicate or clone a speaker's vocal characteristics, and recently developed audio language models can also produce sound effects, singing, or music in addition to speech.

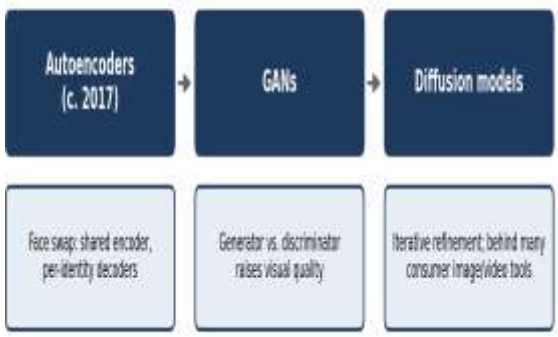


Figure 1: Progression of the main generation techniques behind deepfakes (author's schematic based on [6]).

Figure 1: Evolution of the primary generation techniques for deepfakes (author's schematic based on [6]). The contest between detection and generation is not balanced. The attacker requires only one convincing example, whereas the detector must function across all generation techniques, including those it has never encountered.

Furthermore, social media platforms compress, resize, and screenshot uploaded media, thereby removing many of the subtle cues upon which detectors rely. The initial experiments conducted with FaceForensics++ revealed that humans and manually designed features struggled with compressed video, whereas learned detection systems were considerably more reliable [1]. The laboratory and the real world remain fundamentally different environments.

Table 1: Representative datasets and benchmarks

Dataset	Year	Composition	Main Significance
FaceForensics++	2019	1,000 original videos, four manipulation methods, over 1.8 million manipulated images [1]	Standard training and testing resource
Celeb-DF (v2)	2020	590 real videos and 5,639 fake celebrity videos [2]	More realistic fakes; detector scores fall
DFDC	2020	Large crowd-acted collection; public challenge launched alongside release [30]	Scale and diversity
Celeb-DF++	2025	53,196 generated videos from 22 techniques; 24 detection solutions assessed [3]	Tests generalisation to unseen forgeries
MMTD-Set, DD-VQA, ER-FF++set	2024-2025	Media with accompanying manipulation text descriptions [12][13][14]	Enable explainable detection
ML-ITW, Viewing, AT-ADD, AffectDF	2025-2026	Multilingual, in-the-wild, emotional and all-type audio evaluation sets	Stress-test speech detectors

Source: compiled by the authors from the cited references.

3. Literature Review

3.1 Surveys and Datasets

Several surveys have already provided a comprehensive overview. Nguyen et al. examined deep learning for both generating and detecting deepfakes [4]; Pei et al. compared representative methods in face

swapping, reenactment, talking-face generation, and forgery detection [6]. Wang et al. categorized the field based on reliability, highlighting transferability, interpretability, and robustness as the main issues, and asserted that detector outputs are not yet suitable for legal proceedings [5]. A comprehensive review in 2025 addressed images, video, audio, text, and real-time content [7], while a somewhat more conservative overview in 2022 presented metrics such as accuracy, F1, and AUC. Surveys addressing audio [16] and text also exist. These surveys concur that generators are accelerating their production of deepfakes, while detectors are failing to keep pace, and that standardized datasets are necessary. Most of these surveys were published prior to the emergence of multimodal LLMs, which is the gap this paper seeks to address (refer to Table 1).

3.2 Conventional Detection and Its Generalization Problem

Traditional detection approaches treat the task as a classification problem between real and fake. FaceForensics++ demonstrated that XceptionNet, pre-trained on ImageNet and fine-tuned on the dataset, could identify subtle manipulation cues that are imperceptible to the unaided eye [1]. The creators of Celeb-DF reported that their fakes were significantly more similar to genuine videos and that existing detectors experienced difficulties detecting them [2]. Celeb-DF++ further emphasized this point: all 24 detectors tested demonstrated the lowest performance on Celeb-DF++, and the authors concluded that a universal detector capable of generalizing across various types of forgeries has yet to be developed [3]. An interesting observation emerged from the study: detectors trained on face-swapping data performed poorly on reenactment and talking-face fakes, while detectors developed for the latter two exhibited greater transferability between them, presumably due to their shared motion characteristics.

3.3 General-Purpose LLMs and MLLMs

Jia et al. conducted the first comprehensive study, using a standardized prompt to ask GPT-4V to assess which face images were synthetic [8]. Using only one prompt, the model achieved an AUC of 79.5% on latent-diffusion faces and 77.2% on Style GAN faces (Fig. 2), respectable results for a blind test on a model not trained for the task. More noteworthy was the qualitative distinction between the two types of models: conventional detectors primarily detect signal-level statistical features, whereas multimodal LLMs primarily respond to semantic anomalies and explain them verbally. A subsequent review reported excellent zero-shot performance on image deepfakes and some resilience to post-processing, but also noted models that failed to differentiate between AI-generated and human-created images in specific artistic genres, such as vintage [10]. A 2025 comparison of several Deepfakes and Their Analysis Using Large Language Models Page 5 models, including GPT-4, Gemini 2.5 Pro, and DeepSeek R1, concluded that accuracy generally lags behind specialized detectors, while natural-language reasoning remains a strength [11].

Another report on multimodal facial deepfake detection indicated that alignment, fusion, and prompting continue to be areas requiring further investigation [15].

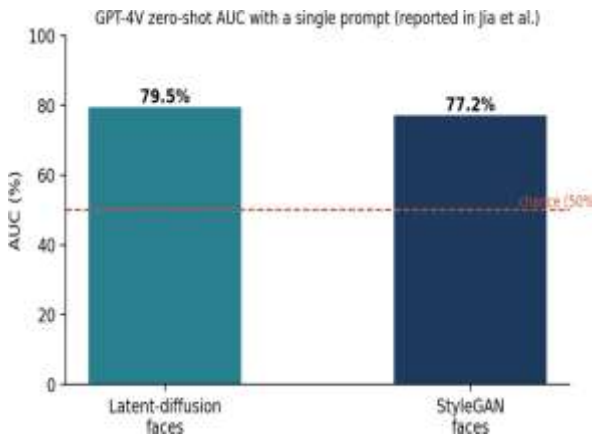


Figure 2: Reported zero-shot AUC of GPT-4V on synthetic face images [8]. Values are taken from the cited study; the dashed line marks chance level.

3.4 Fine-Tuned Explainable Frameworks

A second category of research papers employed MLLMs directly. FakeShield, presented at ICLR 2025, is a framework for image forgery detection and localization: one module evaluates the authenticity of an image and generates a reasoning document, while a second module utilizes this document to identify the manipulated regions [12]. The authors generated a tamper-description dataset (MMTD-Set) and documented improved results compared to previous methodologies. TruthLens relies on two components of its vision stack: one (PaliGemma2) for scene-level comprehension and another (DINOv2) for fine-grained details, enabling it to provide justifications for judgments regarding both manipulated and entirely synthetic faces [13]. For video, EDVD-LLaMA defined a reasoning-based detection task, developed a reasoning-annotated FaceForensics++ dataset, and reported encouraging outcomes on forged videos not included in the training data [14]. A common theme among these studies is that models required to justify their assessments demonstrate greater attention to supporting evidence. These improvements should be viewed critically, as they are derived from the authors' own experiments and lack independent verification.

3.5 Audio and Text

Audio represents the area where the performance of LLMs appears most inconsistent. A 2026 study applied reinforcement learning with frequency-time reasoning to an audio LLM; a 3B-parameter model trained exclusively on speech attained 99.75% accuracy on ASVspoof2019 LA with interpretable rationales, and joint training across all audio types yielded an average accuracy of 90.10% [8].

In contrast, on VoiceWukong, the most effective of twelve conventional detectors (AASIST2) recorded an Equal Error Rate (EER) of 13.50%, while the audio LLM Qwen2-Audio achieved an F1 score of zero on the English set [7]. ML-ITW, encompassing 14 languages and seven platforms, reported significant deterioration in performance under real-world conditions and cautioned that controlled benchmarks might inflate the robustness of these systems [9]. Text generators possess distinct characteristics, such as adding excessive detail or fabricating narratives, which can confuse older detectors. An Arabic study demonstrated that a fine-tuned LLM could detect text authored by ChatGPT and compete with human judgment, emphasizing the necessity of developing evidence specific to low-resource languages [3].

Table 2: Selected results reported in the literature (different metrics and protocols; not for ranking)

Source (as cited in text)	Setting	Reported result
Jia et al. [8]	GPT-4V, zero-shot, single prompt, face images	AUC 79.5% (latent diffusion); 77.2% (StyleGAN)
2026 audio-LLM study [8]	3B audio LLM, RL with frequency-time reasoning, ASVspoof2019 LA	Accuracy 99.75% (speech-only training); 90.10% average (all audio types)
VoiceWukong [7]	12 conventional detectors vs. Qwen2-Audio	AASIST2 best conventional: EER 13.50%; Qwen2-Audio F1 = 0 (English set)
ML-ITW [9]	14 languages, 7 platforms	Considerable degradation under realistic conditions

4. Research Methodology

This study is a comparative, synthesizing literature review. No novel experimentation was conducted. Given the rapid evolution of the topic and the lack of identical experimental conditions across different studies, existing information was collected and aligned rather than retested. Sources were identified through web searches of arXiv and other open-access repositories, the ACL Anthology, ACM, ScienceDirect, USENIX proceedings, and official government communications. Keyword combinations included "deepfake detection survey," "multimodal large language model deepfake," "explainable forgery detection," "audio deepfake LLM benchmark," and "synthetically generated information India." The collected material was compiled in early October 2026 and underwent minimal editing. Only scientific articles and preprints that outlined their methodology were included, along with surveys, prominent benchmark studies, official or legal analyses, and a single documented incident. Marketing materials and content that could not be attributed to authors or institutions were excluded. For each source, the modality, the function of the language model, the performance metrics and associated datasets, the explanations offered, and any evidence regarding robustness were documented. Abstracts, summary sections, and graphical representations of results were primarily utilized instead of re-reading every experimental detail. The authors do not claim credit for the experiments conducted by the cited papers; therefore, the trustworthiness of this review is directly dependent on the reliability of the underlying studies. Original figures are presented as published and may not be based on comparable experimental protocols; consequently, they should not be used for ranking purposes. Some sources were preliminary and may have subsequently been revised, and the review primarily reflects English-language content.

5. Analysis:

The Roles of Large Language Models The literature demonstrates that "LLM" is not a singular methodology.

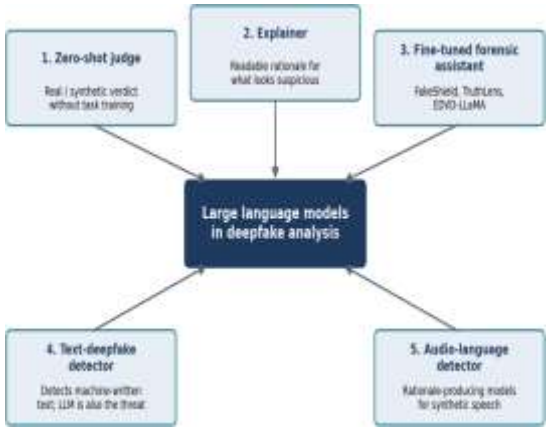


Figure 3: The five roles of language models identified in the reviewed literature.

Zero-shot judge.

A general-purpose model is presented with an image and queried as to whether it is genuine or synthetic. A labeled dataset is

unnecessary, which constitutes a significant advantage. Initial evaluations yielded AUC values in the upper seventies [8], however, on benchmark accuracy specialized detectors consistently outperform general-purpose models [11]. A potential explanation for this disparity is that a zero-shot model seeks characteristics that a professional photographer would identify as anomalous, such as inconsistent lighting or blurred focus on the subject's face, while contemporary generators actively mitigate these common flaws; conversely, a genuine photograph featuring an atypical composition may be misidentified as a fabrication.

Table 3: Conventional detectors compared with LLM-based analysis

Dimension	Conventional trained detectors	LLM / MLLM-based analysis
Evidence used	Very minor distortions and statistical clues	Inconsistencies of meaning and context; knowledge of the world
Explanation	Limited, often heat-maps	Natural-language narrative, possibly citing time or place
Generalisation	Drops on unseen generators [2][3]	Depends less on a particular source, but accuracy can vary [8][10]
Benchmark accuracy	Typically, higher in-domain [11]	Weak zero-shot; good results when fine-tuned [12][8]
Main risk	Over-fitting to dataset quirks	Made-up rationales, prompt dependence, version shift

Explainer

Despite achieving moderate levels of accuracy, LLMs are capable of generating an intelligible explanation detailing why specific elements within a sample appeared unusual and providing a rationale for those observations. This addresses the long-standing black-box nature of detectors [12]. A significant concern associated with this approach is that a generated explanation can sound highly plausible even when it is not supported by empirical evidence; an invented argument can install unwarranted confidence in the validity of a judgment. The authors of EDVD-LLaMA attempt to mitigate this issue through traceable reasoning and quality-controlled annotations [14]. It remains unclear whether this issue has been conclusively resolved by researchers.

Fine-tuned forensic assistant.

With FakeShield, TruthLens, and EDVD-LLaMA, the LLM transitions from speculative assessment to specialized analysis, exhibiting improved cross-method performance and more actionable explanations [12][13][14]. These models are dependent on the linguistic annotations provided to them, and large-scale annotation processes are resource-intensive and inherently subjective.

Text-deepfake detector and generator. LLMs function in both capacities in this domain [2][3], and attributes such as surprisal offer statistical information that enhances interpretation [4]. Detection is increasingly characterized by probabilistic assessments: the distinction between human-authored and machine-authored text is becoming increasingly blurred, to the extent that even the most sophisticated systems cannot reliably differentiate between the two.

Furthermore, imprudent application of an LLM detector in educational contexts poses a risk of unjustly accusing students of plagiarism.

Audio-language detector.

Audio language models are capable of delivering a verdict supported by an interpretable rationale; nevertheless, results are inconsistent: extremely high accuracy on a controlled benchmark in one study [8], but an F1 score of zero for another audio LLM on a distinct benchmark [7], and degraded performance in multilingual, in-the-wild environments [9] (Table 2).

6. Discussion

6.1 Answering the Research Questions

RQ1. General-purpose models exhibit genuine but limited capabilities. They can identify some low-quality fakes and provide explanations; however, their performance is susceptible to variations in model version, prompt, and generator, thus rendering them inadequate for independent decision-making [8][10][11].

RQ2. Language-supervised fine-tuning appears to represent the most promising avenue for enhancing detection accuracy, localization, and explainability.

Furthermore, according to the authors, this approach may potentially lead to improved generalization capabilities [12][13][14]. Prior to placing substantial reliance on these findings, independent third-party validation is required.

RQ3. Image detection is the most advanced domain within deepfake detection, and video detection is progressing favourably. Audio detection remains a major challenge due to its inherent variability. Text detection differs qualitatively from other modalities because the LLM itself serves as the source of the threat. High performance scores in one domain do not necessarily translate to comparable real-world performance [7][9].

RQ4. The potential for hallucinated explanations poses a significant risk, as an AI system can present inaccurate evidence in a convincing manner.

Consequently, deployment strategies should incorporate a human decision-maker within the loop (Section 6.2), transparently disclose error rates, safeguard sensitive facial and vocal data processed by the system, and align with transparency-driven regulatory frameworks (Section 6.4).

6.2 A Layered, Human-Supervised Workflow

Due to the inadequacy of any single verification mechanism, the available evidence suggests integrating multiple layers of verification.

Table 4: Proposed layered workflow for deepfake analysis

Stage	Component	Purpose
1	Provenance and metadata check	Look for platform-generated labels, content credentials or watermarks
2	Specialist signal detector	Run trained artefact, frequency or temporal detectors
3	Multimodal LLM review	Verify semantic plausibility and context; produce a structured rationale
4	Cross-modal consistency	Match voice with lip movements, caption with image, claims with facts
5	Calibrated report	Combine scores, express uncertainty, keep prompt records and model versions
6	Human analyst	A trained reviewer decides; critical cases escalate to a senior reviewer



Figure 4: The proposed layered, human-supervised workflow (conceptual design, not an empirically tested system).

6.3 Lessons from the Arup Incident

Technology constitutes merely one element of defense. An analysis posits that the attack was successful primarily because the employee who observed and trusted the fabricated media authorized a substantial transfer [6]. A flawless detector would never have been consulted. Simple procedural safeguards, such as requiring a second approval for large-value transactions, implementing a delay for confidential matters, and conducting a call-back using a separate contact number, would likely have been more efficacious than any detection system.

While an LLM assistant capable of flagging suspicious communications or guiding users through established procedures could serve as a valuable supplement, it should not supplant existing safeguards.

6.4 Regulation and Ethics

Legislative bodies are beginning to address synthetic media directly. India's Ministry of Electronics and Information Technology issued amendments to the IT Rules, 2021 on 10 February 2026, effective from 20 February 2026. The amended rules define "synthetically generated information" as audio, images, or audio-visual content artificially created or modified to appear authentic [7][8]. Intermediaries are obligated to remove unlawful content within as little as three hours of receiving a valid report. Deepfakes are not prohibited outright; instead, the amended rules prioritize transparency and accountability, and AI-generated text falls outside their purview [8]. Similar trends are evident in the EU AI Act, US state legislation, and Chinese labeling requirements [8]. Mandatory labeling represents an initial step; platforms tasked with distinguishing AI-generated content from human-generated content will be compelled to utilize automated tools, which underscores the importance of accuracy, impartiality, and explainability. Misclassifying legitimate video content or failing to identify malicious deepfakes adversely affects individuals regardless of the specific scenario. From an ethical standpoint, systems that process faces and voices handle sensitive data, and transmitting an individual's photograph to an external model constitutes a privacy-related decision in itself.

Additionally, the "liar's dividend" phenomenon arises: as fraudulent media becomes increasingly prevalent, authentic sources may be erroneously dismissed as fabrications. A responsible system must openly disclose its error rates, must not penalize an individual solely based on a machine's assertion, and must acknowledge the inherent uncertainty present in its responses.

7. Challenges and Limitations

• **Hallucination.** An LLM can be entirely persuasive and yet incorrect; absent mechanisms linking rationales to supporting evidence, the explanation can misinform rather than assist.

• **Sensitivity**. Results are influenced by prompt formulation, repetitive queries, and undisclosed model updates, thereby hindering replicability of research findings [8]

• **Generalization.**
Benchmark performance metrics frequently do not accurately reflect cross-dataset, cross-language, or in-the-wild capabilities [3].

• **Adversaries.** Malicious actors may post-process or paraphrase media, and once a detection feature is publicly known an attacker may develop a system designed to evade it.

• **Cost and scale.**
Deploying a large model for each uploaded item is economically unfeasible, and real-

time monitoring of telephone conversations poses technical challenges.

• **Evidence and language.** Detector output is typically not admissible as conclusive evidence in legal proceedings [5], and the majority of benchmarks are based on English.

• **Nature of this report.** This report does not contribute original research through data collection or experimentation.

8. Future Research Directions

•A unified experimental protocol incorporating state-of-the-art generators, realistic compression algorithms, multiple languages, and diverse audio-visual pairs, facilitating an equitable comparison between LLM-based and traditional detection methods.

• Evaluation of the fidelity of explanations, for instance, verifying whether the features cited by a model are indeed indicative of manipulation.

• Development of hybrid models integrating signal detectors, provenance information, and LLM-driven reasoning, characterized by calibrated uncertainty and resilience to adaptive adversaries.

• Implementation of privacy-preserving deployment strategies, such as small on-device models, and investigation into how individuals interpret machine-generated explanations.

9. Conclusion

Deepfakes have evolved from an experimental curiosity into a pervasive threat to trust, financial integrity, and personal dignity.

While classical detectors continue to offer value, they frequently fail to identify emerging forms of deepfakes and provide minimal justification for their judgments. LLMs introduce semantic comprehension that can be leveraged for explanatory purposes. Based on this review, the capabilities offered by LLMs are genuine but constrained. As zero-shot judges, they facilitate triage but cannot independently render a definitive judgment. When subjected to fine-tuning with language supervision, such models demonstrate acceptable levels of accuracy, localization, and explainability. Performance metrics for audio detection remain inconsistent even in state-of-the-art configurations, and for text detection, the core problem is inherently probabilistic. Both technical findings and the Arup incident underscore the need for a layered approach: provenance verification and labeling, specialized detectors, LLM-driven reasoning, and human verification operating collaboratively, without exclusive reliance on any singular indicator.

For example, India's 2026 amendments establish transparency as a regulatory principle, and the practical implementation of these regulations will ultimately determine their efficacy.

In summary, LLMs are powerful adjuncts in combating deepfakes, but they are not infallible truth-tellers.

References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019.
[2] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207-3216.
[3] Y. Li et al., "Celeb-DF++: A large-scale challenging video DeepFake benchmark for generalizable forensics," arXiv:2507.18015, 2025. [4] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," arXiv:1909.11573.
[5] T. Wang, X. Liao, K. P. Chow, X. Lin and Y. Wang, "Deepfake detection: A comprehensive survey from the reliability perspective," arXiv:2211.10881.
[6] G. Pei et al., "Deepfake generation and detection: A benchmark and survey," arXiv:2403.17881, 2024. [7] "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," review article, ScienceDirect, 2025 (PII S240584402500653X). Deepfakes and Their Analysis Using Large Language Models Page 11
[8] S. Jia et al., "Can ChatGPT detect DeepFakes? A study of using multimodal large language models for media forensics," in Proc. CVPR Workshops (WMF), 2024. [
[9] Z. Yang et al., "The dawn of LMMs: Preliminary explorations with GPT-4V(vision)," arXiv:2309.17421, 2023.
[10] "A review of large language models for multimodal deepfake detection," preprint, ResearchGate, 2025.
[11] "Can LLM detect deepfakes? A multi-modal analysis using state-of-the-art large language models," preprint, TechRxiv, 2025.
[12] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang and J. Zhang, "FakeShield: Explainable image forgery detection and localization via multi-modal large language models," in Int. Conf. Learning Representations (ICLR), 2025.
[13] "TruthLens: Explainable DeepFake detection for face manipulated and fully synthetic data," arXiv:2503.15867, 2025.
[14] H. Sun, C. Cai, H. Zhuang, K. A. Lee, L.-P. Chau and Y. Wang, "EDVD-LLaMA: Explainable deepfake video detection via multimodal large language model reasoning," arXiv:2510.16442, 2025.
[15] "Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges," arXiv:2406.06965.