Deepfakes and Large Language Models: Risks, Defenses, and the Future of Generative AI

Alakananda Mitra *, *Senior Member IEEE*, Saraju P. Mohanty †, *Senior Member IEEE*, Elias Kougianos ‡

***Abstract*—Generative Artificial Intelligence (GenAI) is rapidly changing how digital content is created and consumed. Two widely used GenAI technologies are deepfakes and large language models (LLMs). Deepfakes generate or modify images, videos, audio, and text that imitate real people, while LLMs provide robust language understanding, reasoning, and multimodal coordination. When combined, these technologies significantly increase the realism, speed, and accessibility of synthetic media, raising concerns about misinformation, impersonation, and loss of digital trust. At the same time, the same reasoning capabilities that enable deepfake generation can also be leveraged for detection, verification, and mitigation. This article explores how LLMs strengthen deepfake generation by enabling realistic scripts, coordinated multimodal outputs, and scalable automation. Furthermore, it highlights how LLMs can also be used to fight deepfakes through semantic analysis, cross-modal verification, and provenance-based safeguards. By examining this dual role in an agentic AI setting, the article emphasizes why LLMs are central to both the deepfake problem and its defense.**

***Index Terms*—Generative AI, deepfake, large language models (LLMs), agentic AI**

I. Introduction

Generative Artificial Intelligence (GenAI) has fundamentally changed how digital content is created, shared, and consumed. Deepfakes and large language models (LLMs) are two prominent and increasingly interconnected manifestations of this transformation. Deepfakes are AI-generated or modified images, videos, audio, or text that imitate real people and have become increasingly realistic and accessible [1]. In parallel, LLMs have evolved into multimodal, interaction-supporting, general-purpose systems that can generate coherent language and reason across contexts.

While both technologies have attracted attention independently, their growing convergence enables more scalable, adaptive, and persuasive synthetic media, intensifying challenges for trustworthy and responsible AI, especially in the areas of digital trust, security, and media authenticity [2]. Addressing these challenges requires moving beyond isolated detection tools toward comprehensive, trustworthy AI frameworks built upon pillars of transparency, accountability, and robustness. Such frameworks emphasize media provenance verification and the reliability of AI-driven reasoning.

Relevant governance frameworks include the European Commission's *Ethics Guidelines for Trustworthy AI* [3] and the NIST *AI Risk Management Framework (AI RMF 1.0)* [4], both of which highlight the importance of integrating technical safeguards, human oversight, and lifecycle risk management to ensure responsible AI deployment. While these guidelines provide a foundational strategy for governance, the rapid interplay between deepfakes and LLMs creates unique technical vulnerabilities that necessitate a more granular examination. This convergence, therefore, motivates a thorough investigation of how generative systems jointly influence risk propagation and defense mechanisms within the digital media ecosystem.

> **Key Insight:** The growing realism and accessibility of AI-generated media challenge traditional notions of digital trust. Addressing these challenges requires a system-level understanding of synthetic media and responsible AI design.

The increasing autonomy and accessibility of generative systems make the problem even more challenging. It has become easy to create, modify, and distribute digital media rapidly, often without clear traces of their source or purpose, placing significant strain on traditional content verification and media literacy approaches. Consequently, trustworthy artificial intelligence frameworks are receiving renewed attention to verify the authenticity of synthetic media. Within this landscape, large language models play a central role by mediating the generation, interpretation, and analysis of synthetic content across modalities, as illustrated in Fig. 1.

This article examines deepfakes and large language models from a system-level perspective, emphasizing model behavior and defense strategies rather than implementation-level technical details, to better equip students and practitioners to understand emerging threats and opportunities in synthetic media.

The subsequent sections are organized as follows: deepfakes and their modalities are introduced in Section II. Section III presents deepfake generation as a multimodal pipeline. Section IV examines LLMs as general-purpose, agentic, and multimodal systems, highlighting their im-

*Nebraska Water Center, University of Nebraska-Lincoln, Nebraska, USA
†Dept. of Computer Science and Engineering, University of North Texas, Texas, USA
‡Dept. of Electrical Engineering, University of North Texas, Texas, USA
Corresponding author: Alakananda Mitra (email: amitra6@unl.edu)
Manuscript received Month day 2025; revised Month day, 2025.

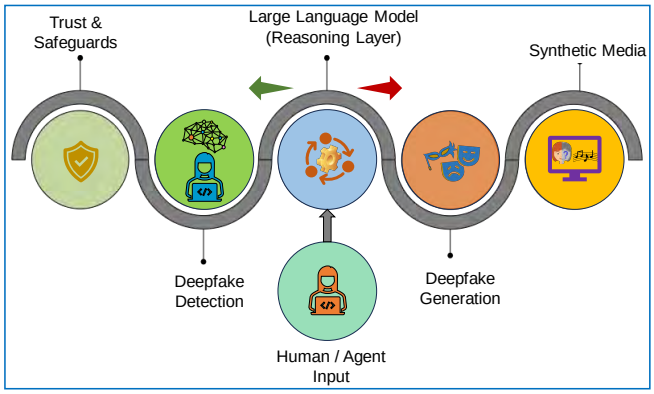

Fig. 1. A high-level conceptual view of the contemporary synthetic media ecosystem, illustrating the coexistence of deepfakes and large language models (LLMs) within modern AI systems. The figure is intended to provide context rather than technical detail.

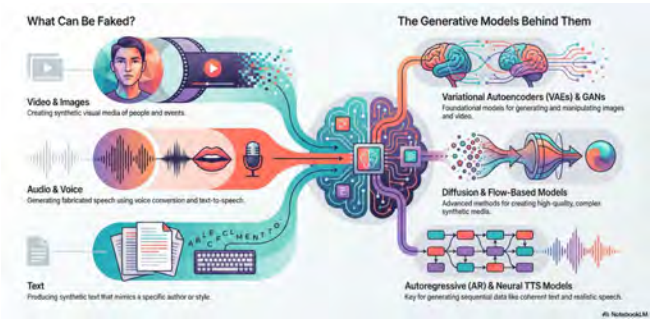

Fig. 2. Conceptual diagram of the core technologies behind deepfake generation, outlining the relationship between synthetic media types and their underlying generative models. Image generated using a generative AI tool based on a conceptual framework and technical prompts provided by the author.

plications for trustworthy media. The relation between LLMs and deepfakes is established in Section V. Section VI outlines some thoughts on efforts to prevent deepfakes. Conclusions are discussed in Section VII. [1]

II. Deepfakes: Threat Model and Modalities

As mentioned earlier, deepfakes are artificially generated or modified media that imitate real people, events, or conversations. Early versions of deepfakes mostly modified existing media, such as through face-swapping or identity transfer. In contrast, modern techniques leverage advanced AI architectures to create completely new content that closely resembles authentic media. Consequently, the term "deepfake" now serves as an umbrella for a multimodal pipeline where images, videos, audio, and text are either altered from the original source or generated from scratch to deceive the viewer. Unlike typical multimedia forgeries that rely on manual editing, these generative approaches produce highly realistic synthetic content, raising serious concerns regarding information integrity [5].

From a security standpoint, deepfakes are effective not only because of their visual realism but also because they deliberately exploit human trust. They can deceive people by convincingly imitating their appearance, voices, linguistic patterns, and behavior, bypassing social and technological safeguards. This makes them particularly harmful in areas like political communication, financial fraud, social engineering, and nonconsensual content creation [2].

Deepfakes can manipulate images, videos, audio, and text (as in Fig. 2). Modern attacks increasingly combine these modalities to create coordinated and persuasive synthetic media. This multimodal nature significantly increases their impact while also complicating detection efforts. For example, a deepfake video may be paired with AI-generated dialogue and a cloned voice to create a coordinated, multimodal deception that is much more convincing than any single modality on its own.

[1]The authors used a genAI tool [8] for technical prose refinement for part of Sections IV and V.

One of the most significant problems with fighting deepfakes is that traditional detection methods are becoming less reliable. Early detection techniques relied on visible artifacts such as unnatural blinking, lighting inconsistencies, or audio distortions. However, advances in generative modeling have significantly reduced these flaws. As a result, deepfake detection is shifting away from artifact-based analysis toward semantic consistency, contextual reasoning, and cross-modal verification [1].

This shift is especially important in the context of large language models (LLMs). Initially, deepfakes were driven mainly by vision and signal-processing models; however, LLMs are now playing an increasingly significant role in determining the content, coherence, and intent of synthetic media.

Hence, understanding deepfakes as a system-level threat, rather than a single model or technique, is essential. This perspective motivates the pipeline-based view of deepfake generation discussed in the next section and supports subsequent analysis of how LLMs influence both the creation and detection of synthetic media.

III. Deepfake Generation as a Multimodal Pipeline

No single model or technique solely drives deepfake generation anymore. Instead, it operates as a multimodal pipeline that integrates visual synthesis, audio generation, and language-based control to produce realistic and persuasive synthetic media. Understanding this pipeline is crucial for recognizing why deepfakes have become increasingly difficult to identify and counteract.

Visual Synthesis Models:

Visual deepfakes rely heavily on deep generative models that learn complex facial representations from large datasets. Early approaches primarily used variational autoencoders (VAEs) to encode facial features into latent representations, as in Figs. 3a and 3b. These representations enabled identity transfer and face swapping by mapping features from one individual onto another.

Autoencoder-based methods remain popular in lightweight deepfake toolkits because they are relatively easy to train and computationally efficient, particularly when the training data is limited.

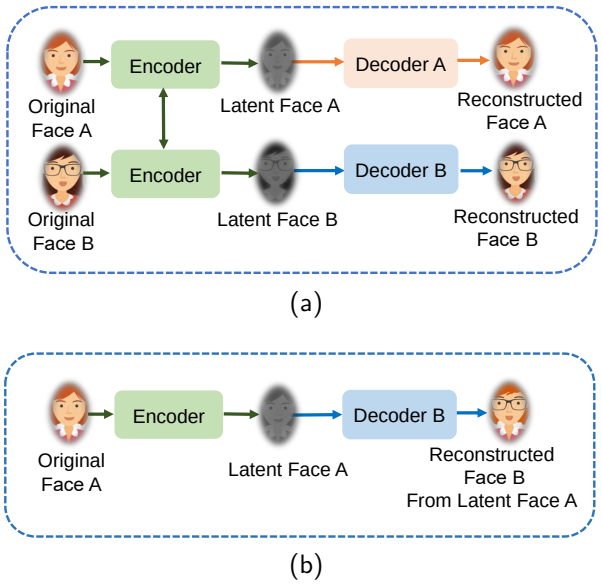

Fig. 3. Deepfake creation using autoencoders: **(a)** autoencoder training and **(b)** deepfake generation [1].

Generative adversarial networks (GANs) significantly advanced visual realism by introducing adversarial training between a generator and a discriminator, as shown in Fig. 4. GAN-based models are effective at synthesizing high-frequency details, such as skin texture, lighting, and fine facial movements, making deepfake videos increasingly indistinguishable from real footage [6]. However, adversarial training can be unstable and prone to mode collapse, motivating the exploration of alternative generative approaches.

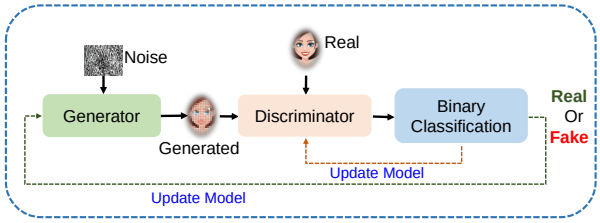

Fig. 4. Training of a GAN to generate deepfakes [1].

More recently, diffusion-based models (e.g., Diffusion Probabilistic Model (DPM)) have emerged as a powerful alternative for image and video synthesis. By progressively denoising random noise into structured data, diffusion models generate high-quality outputs with fewer visible artifacts, such as boundary distortions or temporal inconsistencies [7]. This improvement directly weakens many artifact-based detection methods that rely on visual anomalies.

Autoregressive (AR) models and flow-based generative models also contribute to deepfake creation. Autoregressive models generate images sequentially, learning conditional pixel distributions, while flow-based (FB) models use invertible transformations to model data likelihoods directly. Flow-based approaches are particularly notable for their stable training and accurate reconstruction capabilities, which further complicate detection efforts by minimizing reconstruction errors. Table I summarizes the characteristics of these five generative models.

Key Idea: Deepfakes Are a Pipeline
Modern deepfakes are not produced by a single model. They emerge from a coordinated pipeline combining visual synthesis, audio generation, and language-driven control. This integration explains both their realism and the growing difficulty of detection.

Audio and Voice Synthesis:

Audio and voice synthesis represent critical components of the pipeline. Audio deepfakes are typically generated using neural text-to-speech and voice conversion models, including autoregressive architectures, neural vocoders, and diffusion-based speech synthesis methods. These models enable high-fidelity and expressive speech generation that closely reproduces the target speaker's vocal characteristics. When synchronized with facial movements in video deepfakes, synthetic audio significantly enhances realism and persuasive power. The combination of high-quality voice synthesis with visual manipulation enables impersonation attacks that exploit both auditory and visual trust cues, making detection substantially more challenging.

Language and Control Layer:

Text and language act as the orchestration layer of the deepfake pipeline. Dialogue content, narrative structure, and conversational flow determine how synthetic media is perceived and interpreted. While early deepfake systems relied on manually authored scripts, this role is increasingly automated through large language models.

Toolchains and Accessibility:

A wide range of publicly available tools and platforms has further lowered the barrier to deepfake creation. Tools and services such as FaceApp, DeepFaceLab, Reface, and DeepSwap provide user-friendly interfaces for generating visual deepfakes. Some platforms integrate LLM-driven dialogue generation and real-time conversational avatars, enabling the creation of synthetic media without professional video production expertise. As a result, deepfake generation has become faster, cheaper, and more scalable. This accessibility shifts the threat landscape from isolated, high-effort attacks to widespread misuse by individuals with minimal technical backgrounds.

TABLE I
QUALITATIVE PERFORMANCE COMPARISON OF GENERATIVE MODELS

Generative Models	Sample Quality	Inference Speed	Training Stability	Mode Diversity
Variational Autoencoder	slightly blurred samples	relatively fast because inference involves encoding and decoding	generally stable with proper parameterization	good mode diversity due to their probabilistic latent space
Generative Adversarial Network	high-fidelity and realistic samples	typically fast at inference; adversarial training is computationally demanding	unstable due to the adversarial process	prone to mode collapse, reducing mode diversity
Diffusion Probabilistic Model	high-quality samples, rivaling GANs in image synthesis	slowest due to multiple iterations	relatively stable with careful hyperparameter tuning	highly diverse
Autoregressive Model	high quality but low overall fidelity due to sequential nature	relatively slow because generation is sequential	possible vanishing-gradient issues in long sequences	sequential generation may limit diversity in some cases
Flow-based Models	quality may lag behind GANs and DPMs	often achieve high inference speeds	highly stable due to an invertible architecture and exact likelihood estimation	less pronounced mode diversity

IV. Large Language Models

The evolution of conversational AI can be traced to early systems such as ELIZA (1966), which—although architecturally unrelated to modern transformer-based models—established a foundational milestone in *human–computer interaction* (HCI). ELIZA did not employ probabilistic or neural reasoning; instead, it relied on deterministic pattern-matching and scripted substitution rules. Nevertheless, it revealed the now-canonical *ELIZA effect*, wherein users anthropomorphize digital interfaces and attribute intelligence beyond their actual capabilities. This phenomenon remains highly relevant today, particularly as LLM-driven synthetic media and deepfakes grow increasingly convincing.

Modern large language models extend this historical trajectory. These systems are trained on massive text corpora using deep learning architectures capable of understanding, generating, and reasoning over natural language. Whereas earlier statistical approaches focused primarily on local word prediction, contemporary models exhibit emergent behaviors such as contextual reasoning, instruction following, and multimodal integration.

As illustrated in Fig. 5, the development of conversational AI is categorized into five distinct eras, tracing the transition from early rule-based pattern matching to modern agentic systems. This chronological division highlights the shift from deterministic, scripted interaction toward the probabilistic, multimodal, and reasoning-driven capabilities that define contemporary frontier models. The progression from ELIZA in 1966 to state-of-the-art models through 2025 reflects a steady shift toward more capable, adaptive, and interactive AI systems.

The current model landscape (Fig. 6) is categorized into three functional clusters—reasoning-heavy, vision-language, and efficiency-focused—to demonstrate the trade-offs between cognitive depth, multimodal perception, and deployment speed. This mapping highlights how models like DeepSeek-R1 and the OpenAI o-series (o1/o3) prioritize high-level reasoning for semantic consistency checks, while Qwen2-VL and Gemini 3.0 Pro offer strong spatial reasoning capabilities and native multimodal integration for cross-modal verification. Meanwhile, models such as Phi-4, Llama 3.3, and GPT-4o mini emphasize the computational efficiency required for real-time, on-device deepfake detection without sacrificing the foundational reasoning capabilities needed to identify synthetic artifacts.

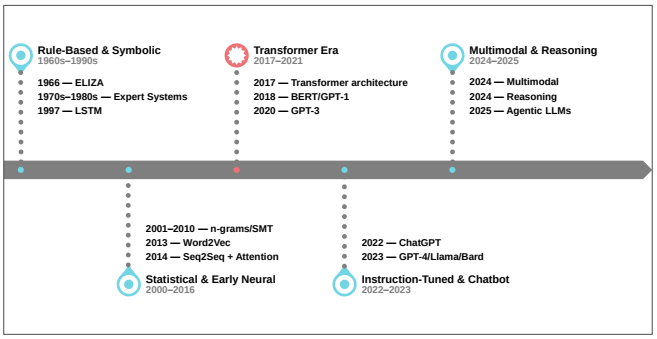

Fig. 5. Chronological development of Large Language Models. Key milestones include the introduction of ELIZA, the 2017 Transformer architecture, and the recent 2024–2025 shift toward agentic AI and "thinking" models.

General-Purpose Models

Recent LLMs are designed as general-purpose models, capable of performing a wide range of tasks, including question answering, summarization, translation, code generation, and decision support. Improvements in model scale, training strategies, and alignment techniques have significantly enhanced their robustness and usability across domains. As a result, LLMs are increasingly deployed beyond traditional natural language processing tasks.

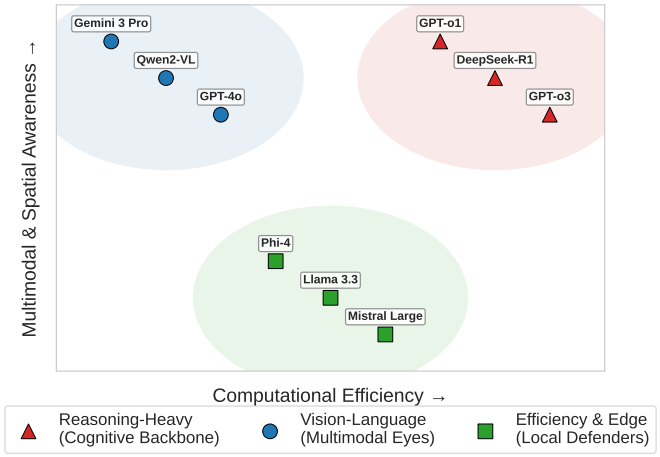

Fig. 6. A conceptual galaxy map of modern large language models, illustrating three dominant design directions: reasoning-intensive models, multimodal vision-language systems, and efficiency-focused architectures. The spatial layout is illustrative and highlights functional clustering rather than quantitative positioning.

Multimodal Input and Output

A defining characteristic of current-generation LLMs is their support for multimodal input and output. Many models can process and generate combinations of text, images, audio, and video, enabling richer human–AI interaction. This multimodal capability allows LLMs to reason across different data representations while maintaining contextual consistency over extended interactions.

Agentic AI

Another significant development is the emergence of agentic behavior in LLM-based systems. In agentic settings, LLMs are integrated with tools, memory, and feedback mechanisms that enable them to plan, reason, and act autonomously toward specified goals. Rather than responding to isolated prompts, these systems can decompose complex tasks, adapt to new information, and iteratively refine their outputs. Such capabilities mark a shift from reactive language models to more autonomous AI systems.

In summary, modern LLMs represent a significant evolution in artificial intelligence, characterized by general-purpose reasoning, multimodal interaction, and emerging autonomy. These properties form the foundation for their growing influence across diverse AI applications, which are examined in subsequent sections. Alongside these advances, concerns related to trustworthy and responsible AI have become increasingly prominent. Issues such as hallucination, bias, misuse, and lack of transparency remain active research challenges.

V. Relationship of LLMs and Deepfakes

LLMs and deepfakes are both products of generative AI. Their relationship presents both promising opportunities and significant risks. LLMs increasingly act as force multipliers within deepfake pipelines, enhancing realism, scalability, and accessibility while simultaneously offering new tools for detection and defense. Table II summarizes the strengths of major LLM families and their relevance to the deepfake context.

Key Insight: LLMs as Orchestrators
Large language models do not generate deepfakes directly. Instead, they act as control and reasoning layers that coordinate language, vision, and audio models, enabling both scalable deepfake generation and adaptive detection and defense.

LLMs Amplify Deepfake Generation

Previously, deepfake systems focused primarily on visual realism and required manual effort to craft dialogue and narrative context. However, the integration of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has fundamentally automated this pipeline. LLMs can automate script generation and coordinate downstream lip-synchronization and speech-generation tools, adapt language and tone across contexts, and maintain conversational coherence. Consequently, a high-level outline may be sufficient to produce credible, context-aware dialogue from models and systems such as OpenAI o3, Gemini 3.1, and ChatGPT, enabling deepfakes to appear natural and situationally appropriate and expanding their potential impact [9]. AV-Deepfake1M is an example of such efforts, in which the authors created $1M$ samples of a ChatGPT-driven audiovisual deepfake dataset. *Deepfake-Eval-2024* shows that LLM-generated scripts significantly enhance realism and coherence, enabling deepfakes to evade both human and automated detection more effectively.

LLMs also enable multilingual and cross-cultural deepfakes, removing language as a constraint for impersonation and misinformation, reducing the barrier to entry, and allowing individuals with limited technical expertise to create convincing deepfakes. Commercial platforms and startups further illustrate this trend by combining LLMs with avatar generation, real-time voice synthesis, and conversational interfaces, enabling high-quality synthetic video production without professional equipment. For example, with OpenAI's new *Sora 2* model, users can now generate lifelike talking heads and personalized avatars through the Cameos feature; startups like Hour One, Synthesia, and Uneeq are integrating large language models (LLMs) for script generation or real-time dialogue, which is then rendered by generative speech and avatar models to produce synthetic video content without traditional filming. Google's new *Chirp 3* can create personalized voices and be used with a deepfake video to produce an audiovisual deepfake. Together, these tools make it possible to generate realistic synthetic audiovisual content with substantially less time, equipment, and specialized expertise than traditional production requires.

From a societal perspective, this convergence raises concerns about misinformation, nonconsensual content,

TABLE II
OVERVIEW OF MAJOR LLM FAMILIES: TECHNICAL STRENGTHS AND DEEPFAKE CONTEXT

Company	Model Family	Key Technical Strength	Deepfake Context (Generation vs. Defense)
OpenAI	GPT-5 / o3	Native multimodal input; chain-of-thought (CoT) reasoning.	Automates real-time interaction patterns and identifies semantic hallucinations in scripts.
Google	Gemini 3.1 Ultra	2M+ context window; SynthID provenance integration.	Enables temporal consistency checks across hours of video; supports provenance analysis through integration with verification mechanisms.
DeepSeek	R1 / V4 Series	Efficiency-oriented MoE; reinforcement learning (RL)-based reasoning.	Supports large-scale, cost-effective monitoring of social media streams for coordinated campaigns.
Meta AI	Llama-4 (Behemoth)	Open-weight innovation with 400B+ parameters; highly customizable.	Powers autonomous adversarial probing and localized, privacy-preserving detection nodes.
Apple	Ferret-UI 2	Fine-grained visual grounding and localized UI interaction.	Detects pixel-level micro-artifacts and regional manipulations on mobile hardware.
Microsoft	Phi-4 Reasoning	Privacy-focused edge reasoning; selective thinking modes.	Defense: Supports on-device analysis that can reduce the need to transmit sensitive media to cloud services.

and the erosion of public trust. As deepfake generation becomes faster and more autonomous, the scale and speed of synthetic media dissemination increasingly outpace manual verification and moderation efforts.

The convergence of LLMs and deepfakes also poses new cybersecurity challenges, such as creating hyper-realistic phishing scams, fraudulent voice calls, and identity impersonation that bypass technical safeguards by exploiting human trust rather than system vulnerabilities. These attacks are challenging to detect because they blend visual authenticity with coherent language and contextual awareness.

LLMs as Enablers of Deepfake Detection and Defense

Despite their role in amplifying deepfake threats, LLMs also play a critical role in detection, verification, and mitigation. Unlike traditional detectors that focus on signal-level artifacts, LLMs can reason about semantic consistency, contextual plausibility, and cross-modal alignment. This allows them to identify contradictions between spoken language, visual behavior, and known facts. Reasoning-heavy models like DeepSeek-R1 perform "logical auditing," identifying semantic glitches and factual contradictions that signal a synthetic script even when visual artifacts are absent.

Recent research shows that multimodal LLMs can assist in media forensics by analyzing relationships across text, audio, and video, complementing conventional deepfake detectors [9]. Modern defense also integrates infrastructure-level protections, such as the *Coalition for Content Provenance and Authenticity* (C2PA) specifications and SynthID watermarking, allowing LLM-based systems to interpret provenance metadata and interact with cryptographic verification mechanisms alongside visual authenticity. LLMs can also support provenance analysis, content authentication, and human-in-the-loop workflows by prioritizing suspicious media for further review.

Industry initiatives further highlight this defensive potential. Efforts such as Meta's Purple Llama project emphasize collaborative red-teaming, safety evaluation, and input–output filtering for generative models, reflecting a growing recognition that LLMs must be integrated with safeguards rather than deployed in isolation.

LLMs as Orchestrators in Synthetic Media Systems

LLMs do not generate deepfakes directly. Instead, they function as control and reasoning layers that coordinate language, vision, and audio models within multimodal systems. Prompting functions allow complex deepfake behaviors to be controlled through natural language as an abstraction layer. By providing high-level intent and structured prompts, LLMs align dialogue, facial expressions, timing, emotional tone, linguistic coherence, contextual relevance, and stylistic uniformity, reducing inconsistencies that earlier detection approaches relied upon. A prominent example is the 2024 video-conference fraud involving an employee of the engineering firm Arup, in which attackers reportedly used digitally manipulated representations of senior personnel to facilitate a fraudulent financial transfer.

Adversarial prompts enable LLMs to rapidly adapt messages across contexts, transforming deepfake systems from passive generators into goal-driven, context-aware interactive personas, thereby increasing their effectiveness in impersonation, fraud, and social engineering. In agentic AI settings, LLMs further enable autonomous interpretation and decision-making across synthetic media workflows, amplifying both capability and scale. Such a system could operate in near real time by leveraging low-latency streaming speech-to-speech models and real-time video-synthesis agents. It could track the progression of a discussion, identify changes, and modify the synthetic persona's emotional reactions and physical actions. Through such integration, deepfakes can function as interactive synthetic identities, making them more dangerous for sophisticated social engineering attacks, high-stakes fraud, and impersonation.

Iterative Feedback and Refinement Loops

LLMs and deepfake tools have a cyclical relationship rather than a linear one. The output from a deepfake system, such as audience responses, adversarial red-teaming reports, or detection outcomes, can be iteratively analyzed and summarized by LLMs to inform system revision. This feedback loop helps enhance the realism of deepfake media by identifying and patching multimodal inconsistencies in near real time. Such interaction distinguishes isolated content generation from coordinated synthetic media systems capable of continuous optimization. This interaction may support closed-loop adversarial refinement, in which generative agents are evaluated by defensive "critic" models and revised to reduce pixel-level and semantic inconsistencies before deployment.

Autonomous / Semi-Autonomous Campaign Architectures

When used together, LLMs and deepfake tools can be viewed as components of larger autonomous or semi-autonomous systems. LLMs excel at decomposing tasks, adapting narratives, and responding to new situations. Deepfake tools, on the other hand, can produce multimodal content at scale. This division of work enables the coordination of substantial volumes of synthetic content with minimal human oversight. Such coordination can accelerate content production and improve consistency across modalities. These architectures may also support multi-agent orchestration, in which a central LLM controller coordinates multiple synthetic agents to maintain consistent but varied narratives across different social media platforms. Importantly, this integration may facilitate sustained, coordinated misuse, raising questions about responsibility and accountability.

Personalization and Trust Exploitation

Fine-grained personalization is possible with LLMs because they can adapt stories to different regional, cultural, or social settings. Deepfake tools enhance this capability by providing perceived authenticity through familiar faces or authoritative voices. These technologies work together to make trust abuse more common by matching personalized messages with believable sounds and images. This has now evolved into automated *linguistic mirroring*, where the LLM analyzes a target's past communications to adopt their specific vocabulary and syntax, weaponizing the *ELIZA effect*—the human tendency to attribute high levels of intelligence and intent to AI. For example, recent digital-insider threats have involved adversarial actors using hyper-personalized synthetic candidates to bypass remote technical interviews and infiltrate cryptocurrency and aerospace firms. Therefore, the persuasive power of synthetic media is not derived solely from how realistic the images are; it also comes from how well the personalization of language and the perceived authenticity of the experience work together.

Agentic AI and the Dual-Use Challenge

The emergence of agentic AI intensifies the relationship between LLMs and deepfakes. In agentic systems, LLMs enable autonomous reasoning, planning, and execution, allowing synthetic media pipelines to operate with minimal human oversight. Such systems can iteratively refine deepfake outputs, respond to feedback, and scale generation rapidly. At the same time, these same agentic capabilities can be leveraged for adaptive detection and monitoring. This has materialized as *near real-time adaptive defense*, where agentic systems continuously analyze live content streams, autonomously update detection heuristics, and deploy countermeasures against evolving attack patterns. Agentic defense systems can continuously analyze content streams, update detection strategies, and respond to evolving attack patterns. This dual-use nature underscores a central challenge: the intelligence that enables scalable deception can also enable scalable defense. For example, reasoning-heavy models such as DeepSeek-R1 and the OpenAI o1 family may support logical auditing by identifying narrative glitches and semantic inconsistencies that complement the capabilities of traditional CNN-based detectors. Similarly, models with massive context windows, such as Gemini 3.1 Pro, allow for temporal grounding and can verify 3D spatial logic and physical consistency across hours of raw video footage. This ensures that the same agentic capabilities used to create convincing fakes are being leveraged to build adaptive, near real-time detection systems.

The correlation between LLMs and deepfakes can be objectively assessed by evaluating the impact of LLM-generated language on realism, scalability, and detectability using the metrics presented in Table III. At the content level, LLMs enhance cross-modal coherence by producing linguistically plausible, contextually aware speech, thereby minimizing lip-sync discrepancies and augmenting mutual information between text and video. At the systemic level, the integration of LLMs significantly improves the generation process and enables the production of several customized deepfake iterations at low cost. From a defensive perspective, detection measures indicate that LLM-enhanced deepfakes degrade both algorithmic effectiveness (as evidenced by reduced AUROC scores) and human evaluation accuracy. This suggests that LLMs may not only enhance deepfake quality but also increase their societal impact. Collectively, these measures provide a framework for evaluating the relationship between LLM capabilities and deepfake-related risks.

VI. Mitigation, Governance, and Societal Response

The growing importance of generative AI has prompted governments, technology companies, researchers, and international organizations to respond through legislation, platform policies, technical safeguards, and transparency requirements. Regulations addressing nonconsensual synthetic media, political misinformation, and disclosure of

TABLE III
QUANTITATIVE METRICS LINKING LLMS AND DEEPFAKES

Dimension	Metric	Quantitative Interpretation
Cross-modal realism	Lip-sync error (ms), phoneme-viseme alignment, text-video mutual information	Assesses audiovisual coherence in deepfake content influenced by LLMs
Narrative quality	Language perplexity, semantic consistency	Captures contextual plausibility and human-likeness of generated speech
Scalability	Time-to-generate, outputs per hour	Quantifies acceleration of deepfake production enabled by LLMs
Personalization	Named-entity accuracy, style similarity	Measures target-specific and context-aware content generation
Detectability	AUROC drop, human error rate	Indicates degradation of automated and human deepfake detection

AI-generated content represent essential steps toward accountability. However, policy responses often lag behind technological advances and remain uneven across regions.

Political communication has been a primary focus, given social media's demonstrated impact on public opinion and democratic processes. In response, major technology platforms have introduced restrictions on the use of generative AI in political advertising, including requirements for disclosure of AI-generated or digitally modified content. Regulatory bodies, particularly in the European Union, have also moved toward mandatory labeling and transparency for AI-powered political advertisements [10]. While these measures address election-related risks, deepfake misuse extends beyond politics. Celebrities and private individuals are increasingly targeted through fraud, impersonation, harassment, and nonconsensual content, underscoring the broader societal impact of synthetic media.

From a technical perspective, academic research continues to propose supervised-learning-based detection methods in multimedia forensics. However, such approaches often struggle to generalize as generative models evolve. This limitation has motivated interest in more adaptive approaches, including reasoning-based analysis and system-level safeguards, which complement but do not replace traditional detectors. To further assist, LLM-assisted methods are being investigated for analyzing manipulated or suspicious real-time content.

Public awareness and media literacy are also critical for countering the effects of synthetic media. As AI-generated content becomes more realistic and immersive, especially in emerging virtual and metaverse environments, distinguishing between authentic and synthetic information will require heightened skepticism and a zero-trust mindset, as emphasized by recent international AI governance frameworks [11].

Ultimately, addressing the challenges posed by deepfakes and generative AI requires a coordinated global response that integrates technology, policy, education, and ethical standards. Such an approach is essential for preserving trust in digital ecosystems while enabling responsible innovation.

VII. Conclusions

The interrelationship between deepfakes and large language models (LLMs) reflects a multifaceted technological paradigm wherein generative systems simultaneously exacerbate and mitigate emergent risks. Individually, each poses challenges to cybersecurity, privacy, and digital trust; together, they amplify these risks by enabling highly realistic, scalable, and accessible synthetic media. Deepfakes constitute a predominant threat vector, undermining epistemic integrity and public trust through the proliferation of synthetic media. Conversely, LLMs play a crucial role in enhancing fluency, coherence, and contextual realism both in generation and defense, requiring careful integration decisions. This duality underscores the necessity for a comprehensive governance framework that integrates technical safeguards, ethical principles, and regulatory oversight.

Future research should focus on developing robust multimodal detection systems, improving explainability in LLM-based verification, and establishing global governance frameworks that integrate technical, legal, and ethical standards. Moreover, socio-technical initiatives like public education and life-cycle risk assessments are crucial for anticipating vulnerabilities and enhancing resilience. Together, these measures may support the responsible evolution of generative technologies by mitigating deepfake-related risks while leveraging LLMs as defensive tools.

Acknowledgments

This article is an updated and revised version of the preprint [12].

Generative AI Declaration

During the preparation of this manuscript, the authors used generative AI-assisted tools to support specific auxiliary tasks. Grammarly was used to improve readability and address grammatical issues; Google Gemini was used for technical prose refinement; and Google NotebookLM was used to assist in generating an illustrative figure. All outputs produced using these tools were carefully reviewed, verified, and edited by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the content presented in this publication.

Read More About It

[1] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," *SN Computer Science*, vol. 2, pp. 1–18, 2021.

[2] M. R. Shoaib, Z. Wang, M. T. Ahvanooey, and J. Zhao, "Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models," 2023. [Online]. Available: https://arxiv.org/abs/2311.17394
[3] European Commission, "Ethics Guidelines for Trustworthy AI," https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, April 8, 2019, accessed on April 14, 2026.
[4] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10, 2023, accessed on April 14, 2026.
[5] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A novel machine learning based method for deepfake video detection in social media," in *Proc. of IEEE International Symposium on Smart Electronic Systems (iSES)*. IEEE, 2020, pp. 91–96.
[6] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "EasyDeep: An IoT friendly robust detection method for GAN generated deepfake images in social media," in *Proc. of the 4th IFIP International Internet of Things Conference*. Springer, 2021, pp. 217–236.
[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[8] Google, "Gemini Deep Research," https://gemini.google.com/app, accessed: Jan. 2026.
[9] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, and S. Lyu, "Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4324–4333.
[10] N. A. Smuha, "Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act)," *International Legal Materials*, pp. 1–148, 2024.
[11] W. Holmes, F. Miao *et al.*, *UNESCO: Guidance for generative AI in education and research*. Unesco Publishing, 2023.
[12] A. Mitra, S. P. Mohanty, and E. Kougianos, "Deepfakes and large language models: Risks, defenses, and the future of generative ai," Feb. 2026. [Online]. Available: http://dx.doi.org/10.22541/au.177023018.89520346/v1

VIII. ABOUT THE AUTHORS

Alakananda Mitra is a Research Assistant Professor in the Nebraska Water Center, Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA. She received her Ph.D. in computer science and engineering from the University of North Texas, Denton, Texas. Her research interests include context-aware intelligence and applied deep learning for computer vision, multimedia analysis, and edge-deployed intelligent systems.

Saraju P. Mohanty is a Professor in the Department of Computer Science and Engineering at the University of North Texas, Denton, TX. He received his Ph.D. in computer science and engineering from the University of South Florida, Tampa, in 2003. His primary research focus centers on "Intelligent Electronic Systems," which has attracted financial support from notable institutions, including the National Science Foundation (NSF), Semiconductor Research Corporation (SRC), the U.S. Air Force, IUSSTF, and Mission Innovation.

Elias Kougianos is a Professor in the Department of Electrical Engineering at the University of North Texas (UNT), Denton, TX. He received his Ph.D. in Electrical Engineering in 1997 from Louisiana State University. Before joining UNT in 2004, he worked at Texas Instruments Inc., Avant! Corp. (now Synopsys), and Cadence Design Systems Inc. His research interests include analog, mixed-signal, and RF IC design and simulation, as well as the development of VLSI architectures for multimedia applications. He is an author of over 200 peer-reviewed journal and conference publications.