Alakananda Mitra[1], Saraju P Mohanty[2], and Elias Kougianos[3]

[1]Nebraska Water Center, University of Nebraska-Lincoln
[2]Department of Computer Science and Engineering, University of North Texas
[3]Department of Electrical Engineering, University of North Texas

February 04, 2026

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

Alakananda Mitra* Saraju P. Mohanty† Elias Kougianos‡

January 26, 2026

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 can generate realistic 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.

Keywords: Deepfakes, Large Language Models, Generative Artificial Intelligence, Synthetic Media, Digital Trust, AI Security, Misinformation, Multimodal AI

1 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 images, videos, audio, or text that imitate real people and have become increasingly realistic and accessible [5]. 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 [8]. Hence, this convergence motivates a thorough investigation of how generative systems jointly influence risk propagation and defense mechanisms within the digital media ecosystem.

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

*Nebraska Water Center, University of Nebraska–Lincoln, USA. ORCID: 0000-0002-8796-4819

†Department of Computer Science and Engineering, University of North Texas, USA. ORCID: 0000-0003-2959-6541

‡Department of Electrical Engineering, University of North Texas, USA. ORCID: 0000-0002-1616-7628

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.

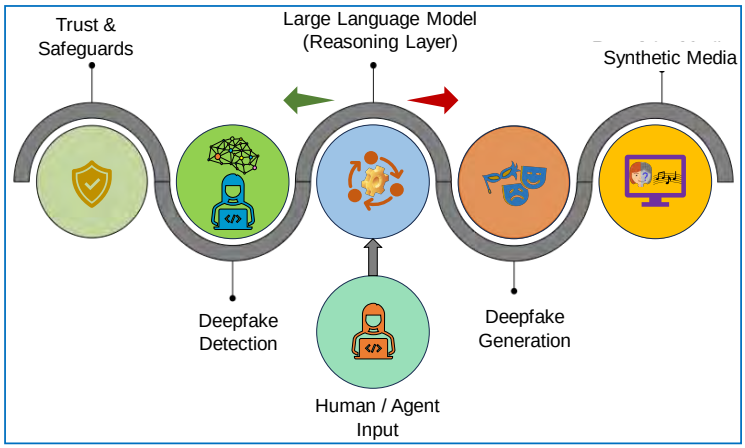

Figure 1: A high-level conceptual view of the contemporary synthetic media ecosystem, illustrating the coexistence of deepfakes and large language models (LLMs) within modern AI systems. The figure is intended to provide context rather than technical detail.

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

The subsequent sections are organized as follows: the concept of deepfake and its modalities are introduced in Section 2. Section 3 presents deepfake generation as a multimodal pipeline. Section 4 examines LLMs as general-purpose, agentic, and multimodal systems, highlighting their implications for trustworthy media. The relation between LLMs and deepfakes is established in Section 5. Section 6 outlines some thoughts on efforts to prevent deepfakes. Conclusions are discussed in Section 7.

2 Deepfakes: Threat Model and Modalities

As mentioned earlier, deepfakes are artificially generated or modified media that mimic actual people, events, or conversations. Unlike typical multimedia forgeries, they use cutting-edge generative models to create content that closely resembles actual media, raising serious concerns about information integrity [3].

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 impersonating their identities, voices, languages, and behaviors, bypassing social and technological safeguards. This makes them particularly harmful in areas like political communication, financial fraud, social engineering, and non-consensual content creation [8].

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 deception type on its own.

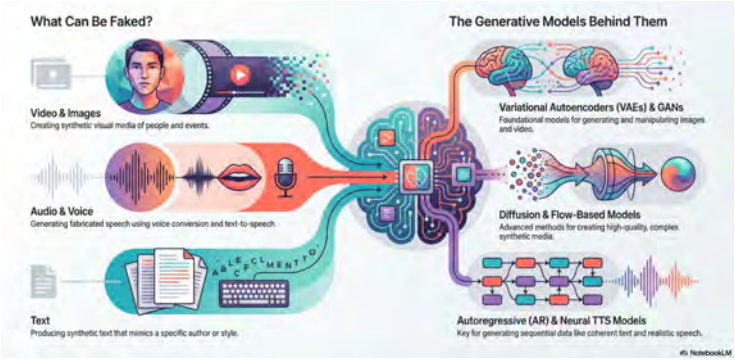

Figure 2: Conceptual diagram of the core technologies behind deepfake generation, outlining the relationship between synthetic media types and their underlying generative models. The photo was generated using an infographic tool in Google NotebookLM studio, where the text was provided by the author.

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 [5].

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.

3 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 training data is limited.

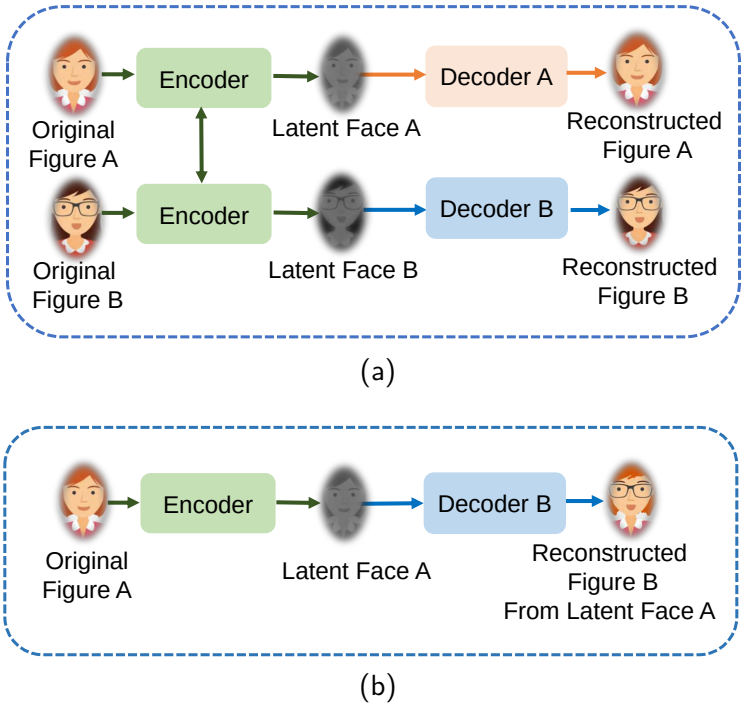

Figure 3: Deepfake Creation using Autoencoder: **(a)** Training of Autoencoders **(b)** Deepfake Creation. [5]

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.

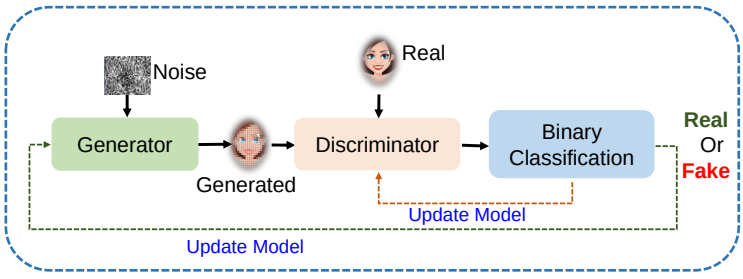

Figure 4: Training of a GAN to Generate Deepfakes [5].

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 [1]. 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. 1 describes the performance of these five generative models.

Table 1: Qualitative Performance Comparison of Generative Models

Generative Models	Sample Quality	Inference Speed	Training Stability	Mode Diversity
Variational Autoencoder	slightly blurrier samples	relatively fast as focused on 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	slower due to their adversarial training	unstable due to 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	Fastest due to their sequential nature	vanishing gradients issue in long sequences	sequential nature limit diversity in some cases
Flow-based Model	quality lag behind GANs and DPMs	often achieve high inference speeds	highly stable due to invertible architecture and, exact likelihood estimation	less pronounced mode diversity

Audio and Voice Synthesis:

Audio deepfakes represent a critical second stage 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 matches the target speaker's identity. 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 have further lowered the barrier to deepfake creation. Applications such as FaceApp, DeepFaceLab, Reface, DeepSwap, and similar services 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 background.

4 Large Language Models

Large language models (LLMs) are deep learning systems trained on large-scale text corpora to understand, generate, and reason over natural language. Early language models focused primarily on statistical word prediction, whereas modern LLMs exhibit emergent capabilities such as contextual reasoning, abstraction, and instruction following. Fig. 5 shows the development timeline of LLMs (only the milestone inventions were noted), starting from the world's first chatbot, Eliza, in 1967 to GPT-5.2, Gemini 3.0 Flash, Mistral Large 3, Grok 4.1, and Claude Opus 4.5 in December 2025 (Fig. 6). These advances have positioned LLMs as foundational components in contemporary artificial intelligence systems.

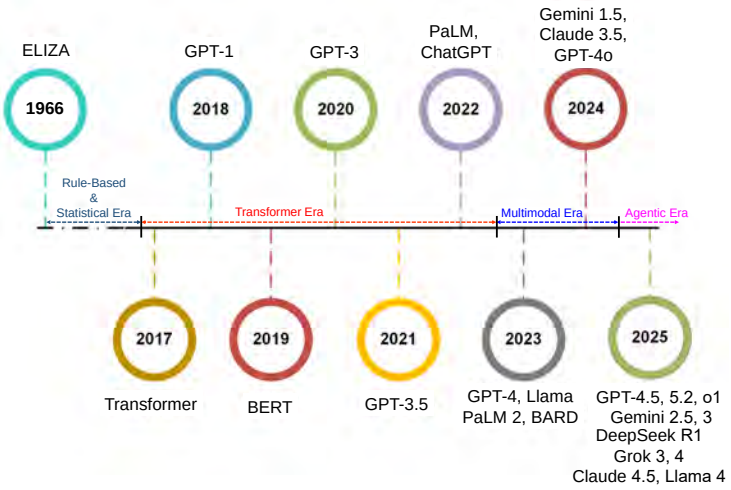

Figure 5: Chronological development of Large Language Models. Key milestones include the introduction of ELIZA, the 2017 Transformer architecture, and the recent 2024–2025 shift toward agentic AI and "thinking" models.

General Purpose Model

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.

Multimodal Input 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.

5 Relationship of LLMs and Deepfakes

LLMs and deepfakes are both products of generative AI. They share both fascinating and alarming relationships. LLMs increasingly act as force multipliers within deepfake pipelines, enhancing realism, scalability, and accessibility while simultaneously offering new tools for detection and defense. Table. 2 summarizes the strengths of major LLM families and their relevance to the deepfake context.

LLM Amplifies Deepfake Generation

Previously, deepfake systems focused primarily on visual realism and required manual effort to craft dialogue and narrative context. If you wanted to create a convincing dialogue for a deepfake video, you had to compose the lines yourself. However, the integration of LLMs fundamentally changes this process. LLMs automate script generation and lip-synchronization, adapt language and tone across contexts, and maintain conversational coherence, enabling deepfakes to appear natural and situationally appropriate. So, you only need to write an outline of the content to obtain a credible dialogue from these models, like ChatGPT, Microsoft's Bing chatbot, or any other text generators, enabling deepfakes to appear natural and situationally appropriate, and expanding their potential impact [4].

LLMs also enable multilingual and cross-cultural deepfakes, removing language as a constraint for impersonation and misinformation. When combined with voice cloning and visual synthesis models, LLM-generated dialogue transforms deepfakes from static media artifacts into interactive

Figure 6: The flagship large language models of six major technology companies: OpenAI (pioneers in conversational AI, focusing on reasoning and complex problem-solving), Google (multimodal models excelling at long document, audio, and video processing), Anthropic (focused on AI safety, ethics, and highly steerable collaboration models), Meta AI (champions of open-source models that are smaller, cost-effective, and efficient), Microsoft (providers of efficient models, including Small Language Models for on-device use), and Apple (a multimodal model specializing in analyzing and answering questions about specific regions within an image). The picture was generated using the prompt "*Generate an iconography showing six major companies (Open AI (GPT and o1 Series), Google (Gemini Family), Anthropic (Claude Family), Meta AI (Llama Series), Microsoft (Phi-3 Series), Apple (Ferret Family)) LLMs*" in Google NotebookLM.

and adaptive synthetic personas. As a result, the barrier to entry has decreased, allowing individuals with limited technical expertise to create convincing deepfakes. AV-Deepfake1M is an example of such efforts, where the authors created $1M$ samples of a ChatGPT-driven audio-visual deepfake dataset. 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 audio-visual deepfake. In a nutshell, these AI tools create videos that do not exist in real life, allowing us to create anything from something, or in some cases, anything from nothing.

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

Table 2: Overview of Major LLM Families (Key Strengths and Deepfake Context)

Company	Model Family	Key Strength	Deepfake Context
OpenAI	GPT-4o, GPT-4.5, o1 family	Strong general-purpose reasoning, multimodal interaction, and step-by-step problem solving	Automates script generation, multimodal coordination, and semantic reasoning for deepfake creation and detection.
Google DeepMind	Gemini 1.5, Gemini 2.x, Gemma 3	Native multimodality with very long-context reasoning and efficient inference	Enables long-context multimodal generation and detection across extended audio-visual sequences.
Anthropic	Claude 3 / 3.5 family	Safety-focused design, controllability, and long-context understanding	Supports controlled reasoning and semantic consistency checks for trustworthy, human-in-the-loop detection.
Meta AI	LLaMA 3.x, Code LLaMA	Open-weight, compute-efficient, and highly customizable	Provides open and customizable components for scalable deepfake generation, detection, and red-teaming.
DeepSeek	DeepSeek-R series	Efficiency-oriented reasoning and coding with reduced computational cost	Supports cost-efficient generation and large-scale monitoring in resource-constrained environments.
Apple	Ferret, Ferret-UI	Region-aware multimodal visual grounding and on-device interaction	Enables fine-grained visual grounding and localized verification of manipulated regions.
Microsoft / xAI	Phi-3 family, Grok-3 / 4	Small, efficient reasoning models and real-world multimodal focus	Supports real-time reasoning and autonomous monitoring in agentic deepfake detection systems.

LLMs as Enablers of 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. Recent research shows that multimodal LLMs can assist in media forensics by analyzing relationships across text, audio, and video, complementing conventional deepfake detectors [4]. 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 [10, 11].

LLM 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. Adversarial prompts let LLMs quickly adapt messages across contexts, turning deepfake systems into goal-driven, flexible, context-aware platforms rather than passive generators, and into interactive synthetic 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.

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 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. Such interaction distinguishes isolated content generation from coordinated synthetic media systems capable of continuous optimization.

Autonomous / Semi-Autonomous Campaign Architectures

When used together, LLMs and deepfake tools can be seen as components of larger, autonomous or semi-autonomous systems. LLMs are good at breaking down tasks, changing stories, and adapting to new situations. Deepfake tools, on the other hand, can produce multimodal content at scale. This division of work enables the coordination of substantial synthetic content with minimal human oversight. This speeds up the process and makes it more consistent, reachable, and fast. Importantly, this integration makes it easier for long-term, planned abuse to occur, raising questions about who is responsible and how they should be held accountable.

Personalization and Trust Exploitation

Fine-grained personalization is possible with LLMs because they can adapt stories to different racial, cultural, or social settings. Deepfake tools enhance this ability by giving fakes greater confidence 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. So, the convincing power of synthetic media doesn't just come from how realistic the images are; it also comes from how well the personalization of language and the reality 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. 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.

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 mentioned in Table. 3. 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 a little cost. From a defensive perspective, detection measures indicate that LLM-enhanced deepfakes degrade both algorithmic effectiveness (as evidenced by lower AUROC scores) and human-evaluation accuracy. This indicates that LLMs not only enhance the quality of deepfakes but also exacerbate their impact on the world. Collectively, these measures establish a method to ascertain the correlation between LLM capabilities and the threat posed by deepfakes.

Table 3: 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 audio-visual 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

6 Mitigation, Governance, and Societal Response

The growing importance of generative AI has prompted governments, technology companies, and researchers to explore mechanisms for mitigating the misuse of synthetic media. Governments, technology companies, and international organizations have begun responding through legislation, platform policies, and transparency requirements. Regulations addressing nonconsensual synthetic media, political misinformation, and disclosure of 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 [9]. 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, advanced algorithms powered by LLMs are now under development to identify manipulated 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 [2].

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.

7 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 lifecycle risk assessments are crucial for anticipating vulnerabilities and enhancing resilience. These solutions will facilitate the responsible evolution of generative technologies, alleviating deepfake dangers while utilizing LLMs as essential defensive instruments.

Acknowledgments

This article is an updated and revised version of the pre-print [7].

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 to expedite the identification of relevant sources; and Google NotebookLM was used to assist in generating illustrative figures. 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.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Wayne Holmes, Fengchun Miao, et al. *UNESCO: Guidance for generative AI in education and research*. Unesco Publishing, 2023.

[3] Alakananda Mitra, Saraju P Mohanty, Peter Corcoran, and Elias 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)(Formerly iNiS)*, pages 91–96. IEEE, 2020.

[4] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei 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*, pages 4324–4333, 2024.

[5] Alakananda Mitra, Saraju P Mohanty, Peter Corcoran, and Elias Kougianos. A machine learning based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, 2:1–18, 2021.

[6] Alakananda Mitra, Saraju P Mohanty, Peter Corcoran, and Elias 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*, pages 217–236. Springer, 2021.

[7] Alakananda Mitra, Saraju P. Mohanty, and Elias Kougianos. The World of Generative AI: Deepfakes and Large Language Models. *arXiv, 2402.04373*, 2024.

[8] Mohamed R. Shoaib, Ze Wang, Milad Taleby Ahvanooey, and Jun Zhao. Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In *2023 International Conference on Computer and Applications (ICCA)*, pages 1–7. Placeholder Organization, 2023.

[9] Nathalie A Smuha. Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act). *International Legal Materials*, pages 1–148, 2024.

[10] Hugo Touvron and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv, 2307.09288*, 2023.

[11] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv, 2310.07704*, 2023.

8 ABOUT THE AUTHORS

Alakananda Mitra is a Research Assistant Professor at the Nebraska Water Center at the 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/RF IC design and simulation and developing VLSI architectures for multimedia applications. He is an author of over 200 peer-reviewed journal and conference publications.