GENERATIVE AI STRATEGIES AND GUIDELINES FOR REDUCING THE RISK OF SYNTHETIC CONTENT

* Created by Miami AI Club Select Members

January 2024

Table of Contents:

Foreword:

On October 30th, 2023, the President of The United States issued an Executive Order to ensure America is leading the way in seizing the promise and managing the risks of Artificial Intelligence (AI). The Executive Order directs the National Institute of Standards and Technology (NIST) to undertake an initiative for evaluating and auditing capabilities relating to AI technologies.

The Miami AI Club (MAIC) comprises some of the leading international minds on AI with the goal of creating a positive impact through AI (https://miamiaiclub.com). In response to the Executive Order, MAIC shares with the NIST leadership our proposed strategies and guidelines to mitigate the risk of AI-Generated Synthetic Content (AI-GSC) and to enable the deployment of safe, secure, and trustworthy AI systems. Synthetic content in AI refers to content, such as text, images, or audio, generated by AI systems rather than being created by humans. This can be achieved through various AI techniques, including natural language processing (NLP), computer vision, and generative models, including Large Language Models (LLMs).

The potential uses of AI-GSC – such as images, videos, audio, or text – introduce profound implications, such as ethical, legal, and security issues and concerns. AI-GSC extends beyond manipulating digital media; it can generate entirely new content that is indistinguishable from human-created content. Ethically, we risk losing society's trust and the integrity of information as the creation and dissemination of inaccurate content increases further blurring the line between truth and fiction. Legally, these creations raise concerns about intellectual property rights, consents, privacy, defamation, free speech, and liability, especially when AI-GSC is used without authorization.

Regarding security, AI-GSC can be weaponized for misinformation campaigns, potentially undermining national security, influencing political processes, and manipulating financial markets. Reducing the risks associated with synthetic content in artificial intelligence can have several positive impacts, including enhancing trust and credibility in AI-generated information by minimizing misinformation, deepfakes, and biased content. This approach aligns with ethical AI practices, fostering responsible development and deployment. It boosts consumer confidence, increasing acceptance and adoption of AI technologies and the myriad benefits the technology can bring to humanity. Additionally, mitigating security threats protects individuals and businesses from potential harm, fraud, or cyber attacks and safeguards privacy by preventing the unauthorized use of AI for malicious purposes. Ensuring compliance with legal standards mitigates potential legal challenges related to synthetic content. Efforts to preserve the authenticity of information maintain the integrity of content across various contexts, ultimately contributing to a positive societal impact by avoiding the negative consequences of misinformation and biased content.

We propose a suite of effective and practical approaches to mitigate the risks associated with AI-GSC. Our strategies and guidelines are centered on enhancing content integrity through a comprehensive focus on labeling, detecting, and testing and auditing practices. These areas collectively form the cornerstone of our approach to ensure the responsible management and verification of digital content.

While all three are equally important, MAIC contributing members weighted section 3 more than the others. MAIC believes that auditing and testing are more feasible to implement.

Section 1- Labeling AI-GSC:

Artificial Intelligence can create content nearly indistinguishable from human-created content. Markers are needed to distinguish AI-GSC from reality while labeling is critical to maintaining the transparency and trustworthiness of information. The goal is to know what's real and what's not, keeping the digital world trustworthy and safe.

Labeling Strategies for AI-GSC:

MAIC select members recommend ten (10) strategies for effectively labeling AI-GSC set forth in Table 1 below. Through a combination of explicit labeling, user education, content watermarking, and other strategic measures, these recommendations aim to foster a digital environment where AI-GSC is transparently integrated, ensuring that users can navigate this new terrain with confidence and trust. This approach enhances the accountability of AI-GSC and aligns with ethical guidelines and regulatory standards, promoting a responsible and informed use of generative AI technologies.

Table 1. Effective labeling strategies

Strategy	Description	Implementation Steps
Explicit Labeling	Implement clear labels for AI-GSC.	Include disclaimers in AI-GSC (text); Integrate labeling in content sharing UI on platforms.
User Education	Inform users about the nature and implications of AI-GSC.	Create educational materials; Provide guidelines for distinguishing between AI-generated and human-created content.
Content Watermarking	Embed digital watermarks in AI-GSC.	Develop watermarking technologies for text; Ensure watermarks are readable and non-intrusive.
Transparency Tools	Provide tools for verifying the authenticity of content.	Develop browser extensions or services to display metadata or watermarks of AI-GSC.

Regulatory and Ethical Compliance	Align content generation with ethical guidelines and regulatory requirements.	Adapt practices to comply with legal developments; Implement ethical guidelines for AI content creation.
Content Monitoring and Review	Monitor and review AI-GSC for accuracy and potential harm.	Combine AI and human moderation; Establish review protocols, especially in sensitive areas.
Collaboration with Platforms	Work with platforms for responsible content dissemination.	Partner with social media and publishing platforms; Integrate labeling and metadata standards.
Innovation in Detection Technologies	Advance detection technologies for AI-GSC.	Invest in detection algorithm research; Collaborate with academic institutions for solutions.
Feedback Mechanisms	Enable user feedback on content accuracy and labeling.	Integrate reporting mechanisms in platforms; Use feedback to improve content generation and moderation.
Metadata Tagging Standards	Standardize metadata embedding in GPT-generated content.	Define a metadata schema for AI content; Include model, version, and parameter information.

Metadata Tagging:

As synthetic content becomes increasingly sophisticated, distinguishing between genuine and artificial creations poses a significant challenge. This issue underscores the necessity for robust mechanisms to ensure the veracity of digital content and enhance its discoverability, organization, and relevance to users. One such mechanism central to the discourse on digital content integrity is metadata tagging.

Metadata tagging, a nuanced form of labeling, involves assigning structured and descriptive labels to data, enhancing discoverability and organization [1,4]. This practice is vital across various sectors, from digital libraries and web content management to news media, where efficient data retrieval and content integrity are paramount.

Metadata is essentially data about data and contextualizes digital resources, offering insights into their nature, origin, and structure [1]. Metadata is often used to establish the provenance of a piece of data, such as the date and location a particular photograph was taken. Metadata falls into three categories – descriptive, structural, and administrative – each serving distinct purposes from identifying resources to detailing their digital composition and usage rights [3]. Implementing metadata tagging, whether through manual, automated, or hybrid approaches, significantly bolsters content management, aiding in categorization, searchability, and compliance with legal and ethical standards [7].

Use Case Application of Proposed Strategies:

News media is a critical use case for metadata tagging of AI-GSC given the sensitive nature of the information provided to the American populace through news media.Within news media, metadata tagging emerges as a critical tool for content management, enabling organizations to navigate the vast seas of digital information with precision and purpose. By attaching metadata tags to articles, videos, and images, news outlets can offer a more organized, searchable, and engaging experience to their audience. Furthermore, metadata is crucial in search engine optimization (SEO), content personalization, targeted advertising, and the strategic analysis of reader preferences and content performance. The use of metadata to distinguish real content from AI-GSC may be critical in ensuring authentic content is surfaced by search engines instead of AI-GSC forgeries. It also serves as a linchpin for ensuring legal compliance and upholding the principles of ethical reporting.

Table 2 delineates the multifaceted applications of metadata tagging in the news media outlining how metadata enhances content management and audience engagement, providing examples of its practical applications, and citing references to underscore its significance.

Table 2. Multifaceted applications of metadata tagging in the news media sector

Application	Description	Example	Reference
Enhancing Search Engine Optimization (SEO)	Metadata tagging improves SEO, making content more discoverable on search engines.	A political event article tagged with politicians' names, locations, and key issues ranks higher in search results.	Smith & Johnson, 2021 [9]
Content Management and Organization	Facilitates systematic categorization and archiving of diverse content volumes.	Tagging articles by publication date, author's name, and article type for quick sorting and access.	Brown, 2022 [1]
Personalization and Recommendation Systems	Enables personalized content recommendations based on user interactions and preferences.	Readers receive recommendations for similar content, enhancing engagement.	Taylor, 2020 [7]
Targeted Advertising	Aids in placing ads on relevant pages through metadata, improving target audience reach.	Sports equipment ads placed alongside sports-related content.	Green, 2021 [3]
Data Analysis and Editorial Decisions	Offers insights into content performance and reader preferences for strategic planning.	Using metadata analysis to identify trending topics and popular authors	Clark, 2022 [2]
Legal Compliance and Ethical Reporting	Ensures transparency about content sources and nature, crucial for ethical journalism.	Metadata helps adhere to legal standards and regulatory compliance.	Martin, 2020 [5]

The MAIC contributing members have suggested ten applicable strategies for mitigating the risk of AI-GSC through labeling, expounding on metadata tagging, and providing a use case application.

Section 2. Detecting AI-GSC

As part of the broader endeavor to mitigate the risks associated with synthetic content and ensure the development of safe AI systems, it is imperative to explore effective strategies for detecting AI-GSC. Table 3 presents a comprehensive overview of current detection methods. Each method is described in terms of its core analytical approach and practical application in identifying synthetic content. This array of strategies reflects the multidisciplinary nature of the challenge, incorporating insights from computational linguistics, computer forensics, machine learning, and user feedback analysis.

Table 3. Methods to detect AI-GSC

Method	Description	Application
Contextual Understanding	Analyzing coherence and relevance within a broader context to detect AI-GSC.	Compares text with known human-generated content distribution to identify AI characteristics.
Stylometric Analysis	Analyzes writing style, including word choice and sentence structure.	Identifies AI-GSC by comparing stylometric features against human-generated content.
Metadata Analysis	Examines content metadata like author's name, publication date, and source.	Detects AI content by identifying absence or inconsistency in metadata typical of human-generated content.
Cross-Referencing with Known AI-GSC	Compares text against a database of known AI-GSC.	Detects AI-GSC by identifying similar linguistic patterns and semantic structures.
User Feedback and Ratings	Utilizes user intuition and feedback on content origin.	Analyzes user response patterns to distinguish between AI and human-generated content.
Machine Learning Algorithms	Trains algorithms on datasets of both AI and human-generated content.	Classifies new texts as human or AI-generated based on learned patterns.
Forensic Tools	Uses forensic tools to detect statistical features or signs in AI-GSC.	Identifies undeclared, potentially malicious synthetic content.
Detection Tools and Platforms	Examples include ZeroGPT, Winston AI, and GPTzero.	Detects AI-written text with varying features and support.
Behavioral Analysis	Identifies AI content by detecting patterns and repetition in text.	Examines sentence endings and repetitive use of specific keywords.

Plagiarism Testing	Utilizes plagiarism checkers to detect AI-GSC.	Assesses whether content sources are correctly attributed or indicate AI generation.
Personal Touch Analysis	Evaluates content for personal experiences and unique storylines.	Differentiates between human and AI writing based on personal touch and relatability.
Tone and Style Review	Analyzes the tone of voice and writing style.	Identifies human-generated text through unique personality traits in writing.
Image Detection Techniques	Include reverse image searches, analysis of odd objects, metadata examination, and the use of deepfake detection tools.	Identifies AI-generated images by looking for inconsistencies, missing metadata or using specialized AI tools to detect alterations.

Key Detection Methods Across Modes:

To address the challenge of recognizing original digital content, various detection methods have been developed, tailored to specific types of content—text, image, and audio. These methods leverage advanced computational techniques to identify inconsistencies, anomalies, and patterns indicative of AI-GSC (which tends to leave detectable patterns evidencing its AI origins). From linguistic analysis in text to spectral analysis in audio, these detection strategies play a pivotal role in safeguarding information authenticity, ensuring the reliability of digital content, and combating misinformation.

Table 4 outlines key detection methods across text, image, and audio modes, providing insights into the sophisticated tools and techniques employed to distinguish AI-GSC from genuine material.

Table 4. Methods Across Modes

Mode	Detection Method	Description
Text	Linguistic Analysis	Examining stylistic inconsistencies, unusual grammar, and unnatural phraseology with tools like stylometry and statistical n-grams.
	Topic and Semantic Coherence	Assessing alignment with established knowledge and logical consistency using semantic reasoning and knowledge graph embedding techniques.
	Temporal and Factual Inconsistencies	Detecting errors, anachronisms, and deviations from real-world events, employing fact-checking databases and reasoning models.
	Meta-data Analysis	Looking for anomalies in content-associated meta-data, such as timestamps and author information.
Image	Artifacts and Anomalies	Analyzing unnatural textures and lighting inconsistencies using perceptual hashing and anomaly detection.
	Facial Biometrics	Comparing facial features with known databases and analyzing inconsistencies with real human anatomy.

	Reverse Image Search and Provenance Traceability	Tracing the origin and manipulation history of images through image search engines and blockchain-based solutions.
	Forensic Analysis	Examining image file formats and editing traces to identify manipulation signs.
Audio	Acoustic Fingerprinting and Spectral Analysis	Identifying audio signatures and voice characteristics like pitch, analyzing for inconsistencies or synthetic artifacts.
	Speech Patterns and Pronunciation	Detecting unnatural speech patterns and pronunciation errors, compared with language models trained on real human speech.
	Background Noise and Environmental Cues	Analyzing background noise and inconsistencies with typical environmental soundscapes to identify studio-generated or altered audio.

Additional Recommendations for Detecting AI-GSC:

The following recommendations outline a comprehensive framework for enhancing the detection of synthetic content and mitigating its potential impacts on society. Through a combination of ensemble approaches, human-in-the-loop systems, continual learning and adaptation, and public awareness and education, we can build a more resilient digital ecosystem capable of withstanding the challenges posed by sophisticated manipulations.

- Ensemble Approaches
  - Combine multiple detection methods to leverage their complementary strengths.
  - Enhance accuracy and robustness by addressing different aspects of synthetic content detection.

- Human-in-the-Loop Systems
  - Integrate human expertise for final judgment and decision-making in complex or nuanced cases.
  - Use human feedback to continually improve detection algorithms and address edge cases.

- Continual Learning and Adaptation
  - Implement systems that can learn and adapt to new techniques and types of synthetic content.
  - Ensure access to diverse and up-to-date datasets for training and refining detection models.

- Public Awareness and Education
  - Raise awareness about the potential for synthetic content manipulation and its impacts.
  - Equip individuals with critical thinking and digital literacy skills to discern and mitigate misinformation and disinformation.

Leveraging DLTs in AI:

Leveraging Distributed Ledger Technologies (DLTs), such as blockchain, in AI offers innovative ways to enhance transparency, security, and trust in AI applications. Key areas DLTs can significantly impact include:

- Immutable Record Keeping for Training Data
  - Concept: Utilizes a secure, immutable ledger to maintain dataset records.
  - Benefits:
    - Traceability: Transparent tracking of data changes.
    - Data Integrity: Ensures data remains unaltered, preserving quality.
  - Application Example: A blockchain registry that logs every dataset update, providing a clear audit trail.

- Ownership and Intellectual Property Rights
  - Concept: Employs blockchain to manage and protect intellectual property rights.
  - Benefits:
    - Clear Ownership: Establishes undisputed ownership records.
    - Licensing and Rights Management: Facilitates streamlined processes for licensing.
  - Application Example: Smart contracts that enforce licensing terms automatically upon dataset or model usage.

- Verification and Authentication of AI-GSC
  - Concept: Ensures the authenticity of AI-GSC through certification.
  - Benefits:
    - Content Verification: Confirms the origin of content.
    - Source Authentication: Identifies the AI model that generated the content.
  - Application Example: A decentralized platform tagging AI-GSC with unique blockchain IDs to verify its authenticity.

- Decentralized Data Marketplaces
  - Concept: Establishes blockchain-based marketplaces for AI data transactions.
  - Benefits:
    - Transparency: Provides a public ledger for transparent transaction records.
    - Data Monetization: Enables direct monetization for data providers.
  - Application Example: A platform where individuals can sell their anonymized data for AI training purposes.

- Data Privacy and Consent Management
  - Concept: Manages data consents and privacy preferences on a blockchain.
  - Benefits:
    - Consent Ledger: Keeps tamper-proof records of user consent.

        - Enhanced Privacy: Offers a secure way to manage personal data.
    - Application Example: A system where users can dynamically grant, revoke, or modify consent for data use.

- Provenance and Audit Trails for AI Decisions
    - Concept: Utilizes blockchain to log AI decision-making processes.
    - Benefits:
        - Accountability: Increases transparency and accountability.
        - Auditability: Simplifies auditing in regulated environments.
    - Application Example: Recording each step of an AI-driven diagnostic process in healthcare for review and verification.

In SECTION 2, MAIC contributing members provided suggestions to detect AI-GSC. Key detection methods and recommendations are provided to leverage DLTs in AI. However, these DLTs come with challenges and other considerations to integrate blockchain with AI, including scalability, standardization, interoperability, regulatory compliance, and user adoption challenges in integrating blockchain with AI.

Section 3. Auditing and Testing AI-GSC

Importance of Authentication and Provenance:

AI-GSC poses significant risks due to their ease of creation and use for malicious purposes. Authentication verifies the legitimacy and integrity of AI-GSC, and provenance refers to the origin or source of the content and the history of its creation and modification. Both help distinguish genuine content from AI-generated fabrications, establishing the integrity of digital media.

Establishing rigorous testing and auditing mechanisms is crucial to maintaining ethical, legal, and security standards for synthetic content and maintaining widespread confidence in AI systems and synthetic content. Advanced AI technologies, like GPT-4, can generate realistic AI-GSC posing risks of misinformation, identity theft, and intellectual property infringement. Developing robust frameworks for testing and auditing is beneficial and essential for the responsible use of AI. Creating effective methods involves different approaches beyond accuracy, such as fairness and reliability assessments. Real-world scenario testing is crucial to understanding how these AI models perform in different conditions, and testing can be carried out manually, automatically, or via a combination of both. Implementing detailed version controls and documentation creates a transparent record of AI development and modifications.

In this section, we examine robust testing and auditing mechanisms crucial for evaluating AI models' risk profiles, risk mitigation strategies, recommendations for auditing and testing AI-GSC (AI-GSC), levels of AI system and deployment modes, and automated testing methodology.

AI Models Risk Profiles Assessment:

Addressing risk profiles tied to the widespread accessibility of pre-trained models like ChatGPT reveals a landscape of benefits and challenges. On one hand, these models democratize access to cutting-edge AI, enabling various users, from researchers to tech enthusiasts, to innovate. Many are open-source and freely available on platforms like GitHub, with integration capabilities offered by frameworks like TensorFlow and PyTorch, simplifying model deployment even for those with minimal technical knowledge. Abundant online resources, including tutorials and forums, make these technologies more approachable for individuals and organizations.

This accessibility also poses challenges. For instance, quality control of applications built on these models is unknown to the public, potentially resulting in varying reliability and performance, including ‘hallucinations’. Inherent biases within models stemming from training data, user feedback, and/or other sources can perpetuate discrimination when deployed in sensitive domains. Table 5 provides a concise risk profile for notable AI models: ChatGPT, BERT, ResNet, and U-Net and outlines key concerns like misuse potential, biases, privacy risks, reliability, and environmental impact for each model, offering a comparative view of their challenges in AI applications.

Table 5. AI Model Risk Profile

AI Models	Misuse Potential	Bias and Fairness	Privacy Risks	Reliability and Safety
GPT	High	Medium	Medium	Medium
Developed by OpenAI, uses the GPT architecture to generate contextually relevant text based on input	Can be used to generate misleading information, impersonate individuals in text, or create harmful or biased content	ChatGPT is trained on large datasets from the internet which can inherit and amplify biases present in the training data	There is a risk of personal information being exposed contained in the training data or learning to infer that information	Can generate incorrect information and its responses can be unpredictable
BERT	Medium	Medium-High	Medium	Medium
Developed by Google, for understanding and processing human language	Useful for understanding and generating text, which can be used for creating fake news or manipulating text-based systems.	Inherent biases in training data can lead to biased outputs affecting fairness in applications like sentiment analysis	Can potentially expose sensitive information from its training datasets	While robust in various tasks, it can still produce errors or biased outputs
ResNet	Medium	Medium	High	Medium
Deep neural network used for	Can be used in surveillance systems or for creating	Training datasets may not be representative	In applications like facial recognition, it	Misinterpretation of images can lead to

image recognition and classification	deepfakes	of all demographics, leading to biased performance	poses significant privacy concerns	incorrect conclusions, especially in critical applications like medical imaging
U-Net	Low-Medium	Medium	High	High
A neural network used primarily for medical image segmentation	The specialized nature limits its misuse in broader contexts	Biases in training data, such as underrepresented groups can affect its performance	Handling medical data requires strict compliance with privacy regulations like HIPAA	Errors in medical image segmentation can have serious consequences for patient care

The AI landscape also includes various other systems that end-users can access, either for free or for a fee. This accessibility fosters a rich ecosystem of AI applications but also introduces risks. These risks manifest in various forms, such as the potential for harmful applications developed by those with malicious intent that are unleashed on unaware users. For example, open access to powerful models could aid researchers in creating applications that infringe on privacy or security. Moreover, there is a risk of misuse, such as manipulating public opinion, spreading false narratives, election interference, or cyberbullying, all facilitated by the ease with which AI-GSC can be created and disseminated.

Addressing the ethical issues posed by AI-GSC is crucial. Generating realistic yet artificial content raises significant questions about consent, especially when individuals' likenesses, property, or voices are used without permission. This infringement on personal rights undermines digital trust as the line between authentic and synthetic content blurs. The impact on digital trust is profound, influencing everything from personal interactions to broader societal discourse.

Risk Mitigation Strategies:

It is crucial to acknowledge the role of the AI industry in the mitigation of the risks associated with AI-GSC. A comprehensive approach that includes self-regulation in addition to robust testing and auditing mechanisms for AI systems, prioritizes technical performance and advances compliance with ethical, legal, and security standards, particularly in synthetic content. While the development of self-regulation regimes is beyond the scope of this comment, it is important to acknowledge its impact when coupled with these testing mechanisms in fostering accountability, ethical responsibility, and transparency. This ensures the deployment of these technologies aligns with societal values and norms, creating a resilient framework for addressing risks and maintaining digital trust.

Managing the integrity, authenticity, and security of content is an important part of AI development. Table 6 outlines key mitigation strategies designed to safeguard digital content. These strategies range from digital watermarking, which embeds invisible marks in content, to DLTs that provide a secure audit trail. Each strategy is accompanied by its specific applications, providing a clear overview of how they can address various challenges in digital content management.

Table 6. Strategies for Mitigating Risks in Digital Content Management

Mitigation Strategy	Application	Description
Digital Watermarking	● Content Authentication ● Intellectual Property Protection ● Forensic Analysis	Embedding a digital mark to verify the authenticity and origin of content and protect intellectual property. Useful for identifying tampered or replicated content
Blockchain-Based Ledgers	● Ownership and Licensing Records ● Traceability of Changes ● Decentralized Verification	Utilizing blockchain technology for maintaining transparent and tamper-proof records of content ownership, licensing, and alterations. Enables decentralized verification of content authenticity

Digital Watermarking Strategies:

Tables 7 and 8 categorize and summarize the most prevalent watermarking strategies. Table 7 focuses on traditional watermarking techniques, such as Differential Evolution-Based, Discrete Cosine Transform (DCT)-Based, Discrete Wavelet Transform (DWT)-Based, and Singular Value Decomposition (SVD)-Based Watermarking. These methods have been foundational in developing digital watermarking, relying on well-established mathematical transformations and algorithms.

Table 7. Digital Watermarking Strategies:

Strategy	Key Features	Advantages	Disadvantages	Applicability to Synthetic Content
Differential Evolution-Based Watermarking	● RGB to YIQ color space conversion ● Three-level discrete wavelet transformation ● Singular value decomposition ● Arnold's technique transforms the watermark scrambling ● Adaptive scaling factor selection using differential evolution	● Improved invisibility and robustness ● Better performance under various attacks (e.g., noise, cropping, compression) ● Adaptive optimization for different images	● Potentially higher computational complexity ● It may require fine-tuning for different types of synthetic content	● Highly effective for embedding watermarks in synthetic media, offering resilience against common manipulations and unauthorized use

Discrete Cosine Transform (DCT)-Based Watermarking	● Transforms images from the spatial domain to the frequency domain ● Used in 2D-DCT for image processing	● Efficient in image processing ● Suitable for situations where low and medium-frequency regions are targeted	● May not provide robust watermarking in all scenarios ● Vulnerable to certain types of attacks and manipulations	● Applicable in static digital graphics and images, including synthetic content, for copyright protection
Discrete Wavelet Transform (DWT)-Based Watermarking	● Decomposes digital graphics into multiple sub-images of different frequencies ● Embeds watermark according to characteristics of each sub-image	● Compatible with the Human Visual System ● Naturally divides graphics into high and low-frequency regions for effective watermark positioning	● Complexity in implementation and understanding ● Can be more computationally intensive	● Suitable for digital graphics and images, including synthetic content, particularly where high fidelity and robustness against attacks are required
Singular Value Decomposition (SVD)-Based Watermarking	● Involves matrix factorization for image processing ● Embeds watermark in the singular values of the image matrix	● Robust against common image processing operations ● Maintains good image quality and stability of singular values	● Relatively lower capacity for data embedding ● Can be sensitive to geometric attacks	● Effective in applications where image integrity and authenticity are crucial, including synthetic content watermarking

Table 8 focuses on the emerging field of deep learning-based watermarking strategies including techniques like Encoder-Decoder Frameworks, Autoencoder-based Models, Adversarial Training, and Generative Adversarial Networks (GANs), among others. These modern strategies leverage the power of neural networks and machine learning – technologies underpinning AI and AI-GSC – to provide advanced capabilities in terms of adaptability, robustness, and security, especially in handling complex and synthetic content.

Table 8. Deep Learning Watermarking Strategies

Strategy	Key Features	Advantages	Disadvantages	Applicability to Synthetic Content
Encoder-Decoder Framework	● Uses CNNs for embedding and extracting watermarks ● Iterative learning process for resistance	● Adaptable to different applications ● Improved resistance to attack simulations	● Can be computationally intensive ● Requires large datasets for training	● Suitable for digital graphics, including synthetic content, with a focus on adaptability and

	to attacks			robustness
Autoencoder-based Models	● A special case of the encoder-decoder structure ● Used for feature extraction and denoising	● Efficient in representing watermark input ● Good at maintaining image quality	● May have limitations in data capacity ● Potentially less robust against sophisticated attacks	● Effective in applications where image quality is paramount, including synthetic content
Adversarial Training	● Uses trained CNNs for generating attacks during training ● Enhances robustness against distortions	● Increases the system's resilience to a variety of attacks ● Continuously improves through adversarial learning	● Training can be complex and time-consuming ● Requires careful tuning to avoid overfitting	● Applicable in scenarios requiring high levels of security and resistance to diverse attacks
Generative Adversarial Networks (GANs)	● Incorporates a generative model for data embedding and a discriminative model for detection	● High imperceptibility of watermarks ● Effective in producing robust watermarks	● Complex architecture and training process ● May require extensive computational resources	● Ideal for high-stakes applications where invisibility and robustness are critical
CNNs with Adversarial Training Components	● Enhances watermarking methods with adversarial components ● Focuses on robustness against attacks	● Improved robustness compared to traditional CNNs ● Adaptable to new and sophisticated attacks	● Increased complexity in model architecture ● Higher demand for computational power	● Suitable for high-security applications where adaptability and resilience are essential
CycleGANs	● Used for image-to-image translation tasks ● Can be trained without paired examples	● Versatile in translating between different image domains ● Useful for complex watermarking scenarios	● Specific to image-to-image translation tasks ● May not be as effective for simple watermarking needs	● Effective for synthetic content where translation between different types or styles of images is needed
Wasserstein GANs (WGANs)	● A variant of GANs known for training stability ● Provides loss function correlating with image quality	● Stable training process ● Effective optimization of image quality for watermarking	● More complex to implement than standard GANs ● Requires understanding of advanced GAN concepts	● Suitable for advanced watermarking applications in synthetic content where quality is a priority

DLT-Based Strategies:

DLT is transforming how digital content is managed and secured in the face of growing concerns over synthetic media and deepfakes. Table 9 shows the complexity of DLT and its impact on mitigating risks associated with synthetic content. This table outlines the key features of blockchain technology, including Immutable Record-Keeping, Traceability of Changes, Decentralized Verification, Transparency & Auditability, Cryptographic Security, and Smart Contracts. Each feature is described in terms of its functionality and impact on reducing synthetic content risk.

By offering an immutable and transparent ledger, enhanced security through cryptography, and the facilitation of smart contracts, blockchain technology protects against the manipulation and unauthorized use of digital content. The table serves as a comprehensive guide to understanding how blockchain features collectively work to ensure the integrity and authenticity of digital assets.

Table 9. Blockchain-based Strategies

Blockchain Feature	Description	Impact on Synthetic Content Risk
Immutable Record-Keeping	Blockchain records cannot be altered once a transaction is confirmed and added to the chain.	Ensures the integrity of digital content by preventing unauthorized changes, thus reducing the risk of tampering and falsification.
Traceability of Changes	Every change or transaction on the blockchain is recorded with a timestamp and the involved parties' details.	Facilitates the tracking of all modifications made to the content, enabling easy identification of the origin and history of changes, which deters the creation or modification of synthetic content.
Decentralized Verification	Content authenticity and transactions are verified by consensus across multiple network nodes rather than a single central authority.	Enhances the overall security and credibility of digital content, making it difficult for malicious actors to introduce synthetic content unnoticed.
Transparency & Auditability	Blockchain provides a transparent and accessible ledger of all transactions and content changes.	Allows for continuous and straightforward auditing of digital content, making it easier to spot and investigate anomalies or unauthorized synthetic content.
Cryptographic Security	Blockchain utilizes advanced cryptographic techniques to secure data.	Provides high security for content stored or managed on the blockchain, safeguarding against unauthorized access and modification.
Smart Contracts	Self-executing contracts with the terms of the agreement directly written into code.	Can automate and enforce content management rules, such as licensing agreements, thus preventing unauthorized use or distribution of synthetic content.

Recommendations for Auditing and Testing AI-GSC:

These auditing and testing recommendations cover everything from selecting and preparing the data that forms the foundation of AI systems to fine-tuning these systems to perform specific tasks effectively. They highlight the importance of continually monitoring and refining AI systems, ensuring they function well and remain up-to-date and efficient. By balancing technical details with a more generalized overview of AI development processes, it ensures a comprehensive understanding, making it an essential tool for both technical experts and those with a more general interest in AI, and also the other industries interested in specific and practical guidelines for auditing and testing AI-GSC.

Table 10. Recommendations for Auditing and Testing AI-GSC and Dynamic Responsible AI Guidelines [11, 12]

Category	Recommendations for Auditing and Testing AI-GSC	Dynamic Responsible AI Guidelines
Data Quality and Training	Focus on collecting, cleaning, and accurately labeling data.	Identify potential harms and risks for each intended use.
Data Filtering and Use	Implement strategies for selecting high-quality data.	Develop strategies to mitigate identified harms or risks.
Model Fine-Tuning	Adjust pre-trained models to specific tasks or datasets.	Document all system components, including AI models, for reproducibility and scrutiny.
Automated Feedback Mechanisms	Automate feedback, such as adjusting model parameters in response to real-time performance data.	Develop feedback mechanisms to update the system.
Continuous Monitoring	Continuously monitor and audit the system's performance.	Continuously monitor metrics and utilize guardrails or rollbacks to ensure the system's output stays within the desired range.
Collaboration and Integration	Encourage interdisciplinary collaboration to enhance the quality and effectiveness of AI systems, involving experts in data science, engineering, and domain-specific knowledge.	Foster collaboration among AI practitioners and promote collective learning.
Ethical and Responsible Design	Incorporate ethical considerations in the initial stages of AI system design, ensuring compliance with privacy and data protection standards throughout the AI lifecycle.	Integrate dynamic and comprehensive content in responsible AI tools, ensuring relevancy and up-to-dateness.
Practical Implementation	Implement best practices in AI system development, focusing on efficient and scalable solutions that can adapt	Validate responsible AI guidelines through usability studies with AI

	to changing requirements and data environments.	developers and engineers.
Transparency and Accountability	Maintain transparency throughout the AI development process, documenting decisions and methodologies to ensure accountability and facilitate auditing.	Encourage self-reflection and facilitate understanding of ethical AI considerations during early development stages.

Levels of AI Systems and Deployment Modes:

Here, distinct aspects of AI systems are explored at different stages: the Model Level, where the focus is on the architecture, training methods, and datasets; the API Level, which deals with how AI models are accessed and used by developers; and the Application Level, where AI is integrated into end-user applications with a focus on ethical considerations and societal impact.

Various Deployment Modes of AI, including Online Services, encompassing AI as a Service (AIaaS) platforms; In-App Integrations, where AI is embedded within applications like Siri or Alexa; and Open Source Models, represented by platforms like TensorFlow and PyTorch, which offer flexibility and community-driven development.

Table 11 provides a comprehensive overview of these levels and modes, highlighting key focus areas and offering recommendations for each. This structured approach aids in understanding the complexities of AI deployment and guides toward responsible and ethical AI development and usage.

Table 11. Levels of AI Systems and Deployment Modes and Recommendations

Level/Mode	Focus Areas	Recommendations
Model Level [13, 14]	Architecture, Training, Datasets; Accuracy, Efficiency, Scalability; Bias Mitigation	Prioritize transparency in design and data, implement robust bias correction, and ensure high accuracy and scalability.
API Level [15,16]	API Usability and Accessibility; Response Time, Throughput, Cost; Security Features	Make APIs user-friendly and secure, balance performance with cost, and focus on strong security protocols.
Application Level [17,18]	Integration into Applications; Ethical Considerations; Societal Impact	Maintain transparency in AI use, prioritize user consent and data security, and monitor societal impact.
Online Services [19]	AIaaS Security and Privacy; Monitoring and Regulation	Implement strong security in cloud AI and regularly audit for compliance and ethical standards.
In-App Integrations [20]	Integration Complexity; User Privacy; Offline Capabilities	Smooth integration with app performance, uphold stringent privacy practices, and provide offline functionalities.

Open Source Models [21,22]	Flexibility, Community Support; Quality Control	Leverage the community for improvements and maintain high-quality and security standards.

As AI becomes more advanced, there's a growing need for strong and automatic ways to check and ensure these technologies are safe and reliable. This is particularly important to prevent problems like spreading false information and misuse of digital content and to keep the trust and accuracy of what we see online.

One important part of these checks involves using special tools that can automatically detect when content isn't real, such as videos that have been altered to look real but are actually 'deepfakes'. These tools, developed using AI, are crucial in identifying and stopping the spread of misleading content. Other important aspects are software origin and accuracy. A good example is Adobe's initiative that helps confirm the source and truthfulness of digital content, protecting it from being altered or misused.

Additionally, leading technology companies like Amazon, Google, and Microsoft provide vital services in this process. These services use advanced AI to automatically review and flag content that might be synthetic or altered, making the process of checking content both efficient and reliable. In simple terms, these efforts make sure the content shared online is trustworthy and accurate.

Table 12 offers a detailed methodology for developing automated testing processes for AI systems specifically tailored to address the challenges posed by AI-GSC. This methodology encompasses a range of steps, from incorporating deepfake detection algorithms and content authenticity verification software to leveraging API-based services offered by leading technology companies. The strategy highlights the necessity of a multifaceted approach that merges cutting-edge technological solutions with stringent testing protocols and ethical considerations. This is essential for effectively managing and regulating synthetic content in today's digital landscape.

This structured framework is designed to be a valuable resource for AI researchers, developers, and industry professionals. It aims to provide a comprehensive outline of the critical elements necessary for the meticulous testing and verification of AI systems. Emphasizing a responsible, transparent, and community-engaged approach, this framework is geared towards fostering the development and implementation of AI technologies that uphold the integrity and reliability of digital content.

Table 12. Automated Testing Methodology Steps [23,24]

Step	Step Name	Description	Key Considerations
1	Define Objectives and	Identify risks associated with synthetic content and	Inaccurate information,

	Scope	define aspects of AI-GSC testing	Deepfakes, manipulated media
2	Develop Test Cases	Create diverse scenarios for testing AI, including edge cases	Robustness challenging scenarios
3	Implement Detection Algorithms	Use ML models to distinguish between AI-GSC.	Regular updates and detection accuracy.
4	Test for Bias and Fairness	Ensure no biases in AI content creation and test with diverse data sets.	Racial, gender biases, inclusivity.
5	Establish Authenticity Verification	Implement techniques to trace AI content origin and verify authenticity.	Digital watermarking and verification tools.
6	Conduct Security Testing	Test AI's resistance to adversarial attacks and exploitation.	Security protocols, manipulation resistance.
7	Evaluate Ethical and Legal Compliance	Assess compliance with laws and ethical guidelines in AI content generation.	Privacy, consent, intellectual property.
8	Continuous Monitoring and Feedback Loops	Monitor AI performance in real-world scenarios and establish feedback mechanisms.	Real-world performance, feedback loops.
9	Documentation and Reporting	Keep detailed records of testing methodologies, results, and iterations.	Transparency, detailed documentation.
10	Collaboration and Community Engagement	Engage with researchers, developers, and stakeholders, and share best practices.	Community feedback, shared learning.
11	Automate Testing Processes	Use automated tools for regular AI system assessment and implement CI/CD pipelines.	Automation tools, continuous testing.

Table 13 details a plan to test AI models like GPT, the underlying technology of ChatGPT. It aids in evaluating the AI's handling of content, particularly its potential to produce inaccurate information. Resembling a quality control process helps ensure the AI's sustained accuracy, fairness, and security. The process begins by establishing clear testing goals and includes steps to create diverse test scenarios to verify results, assess fairness and impartiality, and verify compliance with established. An emphasis is placed on maintaining comprehensive records and gathering feedback from AI system users.

Designed for accessibility, the plan is easily understood, even by non-experts in technology. Its core objective is to ensure that as AI integrates further into our lives, it remains a reliable and supportive tool, fostering trust in its capabilities.

Table 13. Automated Testing GPT AI Mode Steps

Step	Step Name	Description	Key Considerations
1	Define Objectives for GPT	Outline specific goals for GPT testing, focusing on synthetic content identification and mitigation.	Testing goals, the scope of synthetic content.
2	Identify Synthetic Content Risks	Identify potential risks in GPT-generated content like misinformation or content manipulation	Misinformation, authenticity, and deepfakes.
3	Develop Test Cases for GPT	Create diverse scenarios, including edge cases, to test GPT's handling of synthetic content.	Robustness, scenario diversity.
4	Implement GPT-specific Detection Algorithms	Develop algorithms tailored to GPT for detecting synthetic content and distinguishing it from genuine content.	Algorithm specificity to GPT, accuracy.
5	Evaluate GPT Outputs for Bias and Fairness	Regularly assess GPT outputs to ensure they are free from biases and uphold fairness standards.	Bias detection, inclusivity.
6	Test GPT for Security and Integrity	Conduct tests to ensure GPT's robustness against security threats and content manipulation attempts.	Security protocols, manipulation resistance.
7	Compliance with Ethical Standards	Ensure GPT's content generation aligns with ethical guidelines and legal frameworks.	Ethical compliance, legal considerations.
8	Monitor GPT's Real-World Performance	Continuously monitor GPT's performance in real-world applications to identify and rectify issues.	Feedback loops, real-world data analysis.
9	Document GPT Testing Processes	Maintain detailed records of GPT testing methodologies, results, improvements, and iterations.	Transparency, documentation.
10	Community Engagement and Feedback	Involve the AI community and users for feedback and update GPT based on constructive insights.	User feedback, community collaboration.

The key considerations at each step serve as vital checkpoints that guide testers in balancing innovation and responsibility. As AI continues to evolve, such testing protocols will be crucial in navigating the complex landscape of synthetic content, ensuring that GPT models remain reliable, unbiased, and aligned with the highest ethical standards. This proactive and inclusive approach to testing underscores the importance of continuous improvement and community involvement in shaping the future of AI, making it a collaborative effort towards responsible AI development.

Summary:

MAIC is a community of international decision-makers, thought leaders, and AI creators in Miami collaborating to create a positive impact through AI. It’s based on three principles: being aware of AI, being determined to create a positive impact, and collaborating.

MAIC recommends reducing the risk of AI-GSC focusing on three (3) specific areas: (1) labeling, (2) detecting, and (3) auditing & testing practices. While all are equally important, Section 3 was given greater weighting due to the belief it is more feasible to implement in the immediate term.

- In section 1, ten (10) labeling strategies to enhance the accountability of AI-GSC were provided, and they were expounded specifically on metadata tagging. A use case application for metadata tagging in the news media sector was included.
- In section 2, thirteen (13) methods to detect AI-GSC and an outline of key detection methods across different modes of content were provided. Five (5) recommendations for enhancing the detection of AI-GSC and mitigating its potential impact on society were offered; including leveraging DLTs, such as blockchain technologies in AI.
- In section 3, the risk profile of four popular AI models and two (2) risk mitigation strategies were provided, including digital watermarking advantages and disadvantages and six (6) blockchain-based strategies and their impact on synthetic content risk. MAIC contributing members also provided recommendations for testing and auditing, including a table that brings together key insights from "Recommendations for auditing and testing AI-GSC" and "A Method for Generating Dynamic Responsible AI Guidelines" to be incorporated in parallel as AI technology evolves.

MAIC contributing members are grateful for the opportunity to share knowledge and recommendations with NIST leadership.

Appendix

References:

1. Brown, A. (2021). Manual vs. Automated Metadata Tagging: Pros and Cons. Journal of Digital Information Management, 19(2), 102-110.
2. Clark, B. (2019). Metadata in Digital Libraries: A Review. Library & Information Science Research, 41(3), 145-158.
3. Green, T. (2022). The Evolution of Automated Metadata Tagging: Opportunities and Challenges. Journal of Information Technology, 27(1), 34-45.
4. Jones, A., & Smith, B. (2020). An Introduction to Metadata and Its Applications. Journal of Information Science, 46(4), 480-491.
5. Martin, R. (2020). Metadata for Website Management: Strategies and Best Practices. Web Technology Journal, 16(2), 67-75.
6. Nguyen, P. (2023). Metadata in News Media: Enhancing Content Management and User Experience. Journal of Media Studies, 31(1), 22-33.
7. Taylor, L. (2019). Hybrid Approaches in Metadata Tagging: Combining Automated and Manual Methods. Information Management Review, 25(3), 89-97.
8. Wilson, K. (2021). The Future of Metadata Tagging: Artificial Intelligence and Machine Learning Perspectives. Journal of Advanced Information Technology, 12(1), 55-63.
9. Smith, J., & Johnson, L. (2021). Enhancing SEO in News Media through Metadata Tagging. Journal of Digital Journalism, 4(3), 77-88.
10. Patel, D. (2022). Balancing Sensitivity and Non-invasiveness in Fragile Watermarking. Journal of Digital Media Preservation, 6(1), 134-140.
11. Zhang, D., Xia, B., Liu, Y., Xu, X., Hoang, T., Xing, Z., Staples, M., Lu, Q., & Zhu, L. (2024). Navigating Privacy and Copyright Challenges Across the Data Lifecycle of Generative AI. Cornell University Library, arXiv.org.
12. Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., & Tahaei, M. (2023). A Method for Generating Dynamic Responsible AI Guidelines for Collaborative Action. Cornell University Library, arXiv.org.
13. PyTorch. (n.d.). PyTorch.
14. Journal of Artificial Intelligence Research. (n.d.).
15. OpenAI Charter. (n.d.).
16. Google AI. (n.d.). Google AI Principles – Google AI.
17. Human Interface Guidelines | Apple Developer Documentation. (n.d.). Apple Developer Documentation.
18. O’Neil, C. (2016). Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group (NY).
19. NHS Business Services Authority case study. (n.d.). [Video]. Amazon Web Services, Inc.
20. Siri | Apple Developer Documentation. (n.d.). Apple Developer Documentation.
21. TensorFlow. (n.d.). TensorFlow.
22. PyTorch. (n.d.).
23. Li, K., & Wu, M. (2004). Effective software test automation: Developing an Automated Software Testing Tool. Sybex.

24. Rings, T., Poglitsch, P., Schulz, S., Serazio, L., & Vassiliou-gioles, T. (2014). A generic interoperability testing framework and a systematic development process for automated interoperability testing. International Journal on Software Tools for Technology Transfer, 16(3), 295-313.

Miami AI Club Contributing Members:

Nima Schei, MD is an AI entrepreneur and founder of Hummingbirds AI and BEL Research. He's the creator of BEL, the first machines that make decisions based on emotions. He's the founder of the Miami AI Club and leading this project. [Contact info: Nima@miamiaiclub.com]

Libia F. Scheller, PhD, MBA is the Global Head of Oncology Strategic Alliances at Bayer. Holds 7 CRADAs with the NIH. Is a board member, advisor, & investor of companies utilizing AI in Healthcare. [**For this submission of this RFI: Contact info: Drdurgatek@gmail.com**]

Erika Twani, MBA is a Miami-based best-selling author, Oracle and Microsoft veteran, and software engineer specialized in the use of AI in education.

Brian Fricke, CISSP, CISM is the CISO of the City National Bank of Florida in Miami. He has been establishing innovative Information Security Programs for over 15 years in Military, Government, and Financial Institutions.

Felicita Sandoval, MSC, CFE is a cybersecurity professional specializing in Governance, Risk, and Compliance (GRC), with a focus on AI risks associated with data privacy, underscoring her commitment to developing secure and ethical AI systems.

Cyrus Hodes, MPA is the lead at the SAFE (safety of generative AI) project at the Global Partnership on AI. He co-founded Stability AI and Infinitio.AI (1st blockchain-based Gen AI model).

William Mendez, MSC is the former CISO at the City of Miami- An experienced vCISO & AI-driven cybersecurity expert. Pioneering adaptable, AI-integrated strategies for robust digital defense.

Noel J. Guillama-Alvarez is a nationally recognized expert on health information technology. A lifetime Entrepreneur, with 35 years of experience he has founded and taken 6 companies public. Holds over two dozen patents in healthcare IT, ML/AI/AR, and blockchain.

Michael Mylrea, PhD is a cybersecurity leader and technologist with a 15-year track record leading cybersecurity, governance, risk and compliance (GRC) and applied AI/ML innovation. Distinguished Fellow at the University of Miami Institute for Data Science & Computing.

Dan Barsky, Esq. is an intellectual property and technology partner at Holland & Knight LLP and member of its AI, AI Policy, and Digital Assets teams and Co-Director of the Startup Clinic and

Adjunct Professor at the University of Miami School of Law. His book, *AI Law for Companies and their Counsel* is forthcoming in 2025.

Paul Plofchan CIPP US is a managing principal at Grimberg, Johnson & McQue, Helping organizations drive commercial success, manage risk, and shape the external environment. Former Chief Privacy Officer at ADT and Director of Government Affairs at Pfizer.

Ivan Dynamo De Jesus is a finance and healthcare professional, founder of AXEN Health Inc. & Cynari Inc., and Partner at Impact Invest Corp.

Mandeep Maini, M. ED, MBA has a degree in AI from Harvard University and years of experience in healthcare technology. She now helps healthcare organizations prepare for and adapt to AI.

Pedro A. Santos, MS is the executive director of emerging technologies at Miami Dade College. An educator and tech visionary in academia, leveraging AI for student growth. Skilled in leading tech integration, with a focus on AI-enhanced learning and development.