Gen-AI for User Safety: A Survey

Akshar Prabhu Desai *
Google
Mountain View, California, USA
akshard@google.com

Tejasvi Ravi *
Google
Mountain View, California, USA
ravitejasvi@google.com

Mohammad Luqman *
Google
Mountain View, California, USA
moluqman@google.com

Mohit Sharma *
Google
Mountain View, California, USA
mohitzsh@google.com

Pranjul Yadav
Google
Mountain View, California, USA
pranjulyadav@google.com

Nithya Kota
Google
Mountain View, California, USA
nithyakota@google.com

***Abstract*—Machine Learning and data mining techniques (i.e. supervised and unsupervised techniques) are used across domains to detect user safety violations. Examples include classifiers used to detect whether an email is spam or a web-page is requesting bank login information. However, existing ML/DM classifiers are limited in their ability to understand natural languages w.r.t the context and nuances. The aforementioned challenges are overcome with the arrival of Gen-AI techniques, along with their inherent ability w.r.t translation between languages, fine-tuning between various tasks and domains.**

In this manuscript, we provide a comprehensive overview of the various work done while using Gen-AI techniques w.r.t user safety. In particular, we first provide the various domains (e.g. phishing, malware, content moderation, counterfeit, physical safety) across which Gen-AI techniques have been applied. Next, we provide how Gen-AI techniques can be used in conjunction with various data modalities i.e. text, images, videos, audio, executable binaries to detect violations of user-safety. Further, also provide an overview of how Gen-AI techniques can be used in an adversarial setting. We believe that this work represents the first summarization of Gen-AI techniques for user-safety.

***Index Terms*—GenAI, Machine Learning, User Safety, Trust & Safety**

I. Introduction

Machine Learning and data mining techniques are used across domains to detect violation of user safety. Examples include classifiers used to detect whether an email is spam or a web-page is requesting bank login information. However, existing ML/DM classifiers are limited in their ability to understand natural languages w.r.t the context and nuances. The aforementioned challenges are overcome with the arrival of Gen-AI techniques, along with there inherent ability w.r.t translation between languages, fine-tuning between various tasks and domains.

In this manuscript, we provide an overview of the various domains across which user safety can be violated. In particular, we provide more information on how Generative Artificial Intelligence (Gen-AI) techniques can be used towards reduction of egregious abuse (e.g. phishing, malware, anomaly detection, counterfeit, fraud prevention), misinformation and disinformation (e.g. fake news , deepfake detection), increase in content moderation, awareness about mental health (e.g. cyber-bullying prevention, crisis support) and towards robust physical safety (e.g. accessibility, autonomous systems).

Further, we discuss how Gen-AI techniques can be used across various data modalities. In particular, we present how Gen-AI techniques can be used to detect user-safety violations in text and rather outperform all previous techniques w.r.t NLP tasks such as entity recognition, question answering and sentiment analysis. Further, Gen-AI techniques with their inherent ability to parse and understand images provides an easy mechanism to detect image manipulation, deepfake detection. The advantages of Gen-AI techniques goes beyond text and images to other data modalities such as videos, audio and executable binaries.

We also discuss how Gen-AI techniques can be used in an adversarial setting. In particular, we present how these techniques can be used to attack at scale (e.g. mass spam). Further, the attacks become more intelligent as these techniques can target humans more effectively and engage with humans with a similar cognitive capacity. Gen-AI techniques along with reinforcement learning techniques can be used to create more sophisticated attacks while using feedback from the last failure. Further, Gen-AI techniques can also make these attacks look very personalized (e.g. deep-fakes) and second order effects (e.g. Gen-AI imitating human sounding text).

The organization of the paper is as follows. In section 2, we provide a comprehensive overview of the various user safety domains, where Gen-AI techniques can be applied. In section-3, we discuss the various data modalities across which Gen-AI techniques can be utilized to protect user safety violation. In Section-4, we present, how Gen-AI techniques can also be used in an adversarial setting. In Section-5, we present our opinion on what does the future look like for Gen-AI techniques. Finally, in section-6, we conclude this manuscript.

II. User Safety Domains

Gen-AI techniques can significantly enhance user safety in both digital and physical environments. It can proactively address risks, offer timely assistance, and empower individuals with personalized tools. Within digital realm, it can detect

* These authors contributed equally.

fraud, flag instances of harmful content, cyberbullying, and predatory behavior. In the physical realm, Gen-AI techniques can help improve accessibility via wearables, contribute to mental and physical health by providing timely access to online resources and perform crisis intervention if necessary. Below we explore and detail the research that enhance user safety in the two realms.

A. Digital Realm

1) Online Threat Protection: **Phishing** is a type of socially engineered cyber-attack where bad actors try to trick users into giving up their personal information by pretending to be a trustworthy source either via email and phishing websites [43]. The work by Koide et al. [58] show that Gen-AI techniques lowers the barrier to deploy systems that detect, mitigate existing and deter new phishing attacks by utilizing their broad knowledge base, multi-modal and multi-lingual capabilities, which otherwise would require multiple different classifiers [67]. Further, Koide et al. [59] present Gen-AI techniques can provide user with detailed reasoning about why certain email or website is highlighted for phishing. Ai et al. [4] propose that when Gen-AI is augmented with advanced strategies like Retrieval-Augmented-Generation (RAG) it can detect sophisticated phishing attacks that involve longer multi-turn interaction between bad actor and unsuspecting users.

Malware are malicious software, that include viruses, worms, and Trojan horses, deliberately designed to compromise computer systems, servers, or networks. These attacks usually lead to data breaches (e.g., via spyware), system damage, or unauthorized control of devices (e.g., through adware and ransomware). Gen-AI techniques offer the potential to enhance existing strategies for malware identification and alert generation that have traditionally employed machine learning and deep learning models [11] [71]. Ferrag et al. [28] show that with advanced techniques, such as Half-Quadratic Quantization (HQQ), Direct Preference Optimization (DPO), GPT-Generated Unified Format (GGUF), Quantized Low-Rank Adapters (QLoRA), and Retrieval-Augmented Generation (RAG), Gen-AI can be leveraged for more effective threat detection and response. In AppPoet, Zhao et al. [120] demonstrate a system that uses multi-view prompt engineering to detect and produce a detailed diagnostic report for android malware. Similarly Wang et al. [104] in their work ShieldGPT, show that via prompt engineering, Gen-AI has the potential to defend against Distributed Denial of Service (DDoS) attacks and provide comprehensible explanation and detailed mitigation instructions specific to an attack.

2) Misinformation Detection: **Fake News** is false or misleading information presented as news. The proliferation of misinformation across social media platforms and news outlets presents a significant threat in the digital age. The sheer volume of online content renders manual fact-checking impractical and this is where Gen-AI can help. Zhang et al. [117] propose a hierarchical prompting method that outperforms state of the art fully-supervised approach for news claim verification. The work by Yue et al. [113] proposes a retrieval augmented response generation system to combat online misinformation and generates counter-misinformation responses based on the scientific evidences. Furthermore, work by Xuan et al. [108] show how Gen-AI can utilize external knowledge bases for information verification.

Gen-AI are proving effective in detecting fake news, both independently and in conjunction with specialized Small Language Models (SLMs) [46]. Notably, Gen-AI can identify misinformation generated not only by humans but also by automated systems [96]. Their capabilities extend to multimodal misinformation encompassing text, images, and videos [103].

Deepfakes are synthetic media where an entity in an existing image or video is replaced with another entity's (usually a person) likeness using artificial intelligence techniques. Recent advances in Gen-AI has enabled creation of extremely high fidelity personalized content like images, audio and video. At the same time, several techniques has been developed to detect such deepfakes across different modalities [17] [114] [33].

3) Content Moderation: Content moderation is critical to maintaining the integrity of online platforms and to keep the members safe from harmful content, misinformation, and disinformation, and to comply with legal and policy standards. Content moderation is a growing challenge as the platforms scale and sheer volume of content to be moderated can't be handled by humans alone. Diversity of content in terms of language and modality and evolving nature of harmful content itself add an additional challenge.

In-context learning [27] ability of Gen-AI techniques have been used to encode online platform policy in a prompt to detect whether content violates it or not [60]. Multi-modal reasoning capability of Gen-AI now allows us to capture the necessary context spread across text, images, video (short and long form) components of the context to make a judgement on policy violation [42] [2] [3] [13]. This can be extended to create more effective **age-appropriate content filters** to ensure users are not exposed to inappropriate content.

B. Physical Realm

GenAI-powered chatbots can aid in **crisis support** by providing rapid targeted information, enhancing communication, offering emotional support, and improving preparedness. Often during emergencies, emergency support systems are overwhelmed. Advanced frameworks that leverage Gen-AI can be utilized help the support systems by understanding user needs and creating workflows for government agencies [76], [87]. The work by Otal et al. [75], [76], explores using fine tuned models like LLama2 to assist users with simple instructions while informing authorities with summarized and accurate information. Gen-AI techniques can assist with **accessibility**, enabling safer navigation in the world for visually impaired people, for example, in VisionGPT [102], Wang et al. showcase a system that takes in real time video via camera captured frames and provides a concise audio description to enable safe navigation. GenAI can help in data augmentation and synthesis to bridge gaps in existing simulation data and thereby improving autonomous vehicles capabilities [119].

Counterfeit goods are fake products designed to look genuine, like branded items. Production and distribution of counterfeit goods infringe on intellectual property rights which can cause serious damage to consumer health and safety (Counterfeit medicines, cosmetics, or safety equipment) and brand reputation, eventually resulting in losses for legitimate businesses. According to the National Crime Prevention Council around $2 trillion worth of counterfeit products are sold to consumers annually [72]. At the time of writing, there isn't much research around using Gen-AI techniques for counterfeit detection, however, there are relevant work in identifying counterfeits with Generative Adversarial Networks (GANs). The generative component mirrors the counterfeiter's role, while the discriminator functions as the detective, identifying and rejecting fraudulent outputs. Some examples include, a combination of external attention GAN with deep convolutional neural networks (CNNs) developed by Peng et al. [80] to identify counterfeit luxury handbags and GANs for credit card fraud detection as shown by Wang et al [101].

1) Mental Health and Well-being: Mental health and well-being is a topic that is central to user safety and GenAI can be applied to detect, intervene, prevent and provide timely support to ensure users mental well being. VITA [94], a multi-modal Gen-AI based system for mental well being, allows robotic coaches to autonomously adapt to the coachee's behaviours from features like facial valence and speech duration. Using these signals, it delivers adaptable coaching exercises to promote mental well being. Along with useful intervention activity recommendations, Gen-AI based agents that are anthromorphic are able to foster relational warmth and can prove to be more effective [107].

Gen-AI can be trained to detect and flag **cyberbullying** instances in online interactions, which often results in significant psychological distress for the victims, allowing for faster response and support for victims. Vanpech et al. [98] proposed a system to identify cyberbullying via images by feeding the image as input to GPT-4 to generate description metadata, which was then provided to a custom trained Gen-AI model that classified the images to detect if it's used for bullying. Gen-AI can also be used to augment training data to improve existing classifiers. For example, in the work by Jahan et al. [51] GPT-3 based data augmentation showed 0.8% improvement in classification F1 score for hate speech detection tasks.

Predatory behavior detection is a critical research area for social media platforms to ensure user safety, especially for vulnerable populations. GenAI can analyze patterns in text and images to identify grooming behaviors or attempts to solicit explicit content from minors. This can help platforms intervene proactively and protect users from online predators. While out of the box foundational models are already helpful in detecting such interactions, fine tuned models perform better, as shown by Nguyen et al. with a LLama 2 model that was fine tuned using LoRA. [73].

III. Data Modalities

Machine learning has long played a crucial role in enhancing user safety. However, various machine learning techniques have traditionally operated in isolation, handling different data types separately. Advances in Gen-AI now allow for a more holistic analysis of different data modalities. This section explores how Gen-AI impacts user safety across these different data modalities.

A. Text

Existing ML techniques are limited in their ability to detect harmful content across languages (e.g. less spoken languages). Further, separate classifiers are trained for single tasks, thereby increasing complexity of technical stack.

Large language models, such as GPT-4 , Gemini and LLaMA have demonstrated outstanding performance across downstream NLP (Natural Language Processing) tasks (e.g. text classification, named entity recognition, translation, question answering and sentiment analysis) [118]. The advantage that Gen-AI bring is the vast inbuilt context from pre-training enabling transformer based models like RoBERTa perform well even on tasks that were difficult earlier, such as sarcasm detection [82]. In particular for user safety, pretrained transformer models achieve remarkable performance in hate-speech detection, detecting spam, fake news and fake reviews. Further, Keyan et al. demonstrated that well crafted reasoning prompt can effectively capture the context of hate speech by fully utilizing the knowledge base in Gen-AI models, significantly outperforming existing techniques [39]. Using retrieval augmented generation (RAG), these models are able to perform fact verification [62], and are also able to detect fake news written in high quality journalistic style [106].

Further, Gen-AI techniques are also able to overcome content moderation to low resource languages as well as to multilingual text. In particular, multilingual-BERT and XLM-RoBERTa, each of which have been pre-trained on 100+ languages have been used to perform well on hate speech and hostility detection with multilingual input [109] [90]. Further, they are also being used to generate annotated data for training models for low resource languages. Furthermore, annotated data is found to be on par with human annotators and can be done at a fraction of time and cost [56].

B. Images

Following the breakthroughs in large language models, large multimodal models (LMMs) such as GPT-4v [111] extend these capabilities to other data modalities, with images being a particular focus. User safety in the domain of images involves detecting harmful images, images that violate platform policies, flagging sensitive media and/or assisting users in detecting machine generate image from real images.

1) DeepFake detection: Easy availability of image generation tools such as Midjourney, StableDiffusion [88], Dall-E [12] and others have led to proliferation of fabricated images online. These models generate hyper realistic images that are not easily distinguished from real images. Existing techniques for detecting deepfake images are mostly formulated as binary classification problems and fall into three major categories - identifying inconsistencies exhibited in the physical/physiological aspects in the DeepFake images, methods using signal-level artifacts introduced during the synthesis process, or directly training a classifier on real and DeepFake samples. Some of these techniques require large amounts of labeled training data. Another shortcoming is that classifiers trained to detect images generated from one class of generative models (e.g., GAN) fail to generalise on images generated from other class of generative models (e.g., diffusion models). [74]. This is because the artifacts produced by one breed of generative models, which the classifier learns to identify, are different from artifacts produced by a different class of generative models. The same techniques that have resulted in the success of image generation have also been employed to detect deepfake images. Most recently, using a pre-trained CLIP-ViT model to learn image features followed by a classifier to detect fake images sets new state of the art and also generalizes across different breeds of generative models [74].

2) Harmful images detection: Recent advances in vision-language models have significantly improved the ability to detect harmful images, making them useful for content moderation [40] [99]. Models like VinVL, which leverage transformers and attention mechanisms, can capture complex relationships between visual elements and generate accurate, contextually relevant captions [116]. Furthermore, large-scale multimodal pre-training on massive datasets of images and text, such as CLIP (Contrastive Language-Image Pre-training), has enhanced the models' ability to connect visual and textual information [85]. This is crucial for identifying harmful images, especially those containing hate speech or offensive symbols, which often rely on the interplay between visual and textual elements.

Cutting-edge vision language models like GPT-4v and multimodal Gen-AI techniques like GPT-4 and Gemini have shown remarkable understanding of complex concepts in images. PaLI-X, a vision language model trained using trained using latest techniques of instruction-tuning and distillation, outperforms all prior VLMs achieving state-of-the-art performance across complex vision tasks including the Hateful Memes Challenge, which seeks to identify multimodal hate speech [48]. A comparative study of different image safety classifiers done in [84] found that Vision Language Models (VLMs) can identify a wider range of unsafe content, with GPT-4v being the top performing model for this use case.

3) Safety applications: VLM's nuanced understanding of a scene has applications in defect recognition, safety equipment recognition [111]. These models are being used to detect safe pedestrian crossing, adherence to workplace safety guidelines, and evaluation of construction site safety among others [50] [97] [19].

C. Videos

Consumption of video content has been on steady increase especially due to social media and streaming platforms [20]. User safety issues in videos could be harmful content, age inappropriate content, violent content, copyright violations and deepfakes [1] [115] [52].Technology companies have built sophisticated solutions to understand video content and protect users from these threats [53]

Traditional machine learning approaches to user safety have relied upon building specialized models for each kind of threat. These models are built and trained internally by organizations on limited data that is available to them. Gen-AI solves this problem much easily by providing complex and large models that are trained on signifiancly more amount of data and also can achieve multiple objectives with single model.

Gen-AI for user safety in videos can be classified across following subcategories of the video content.

1) Human generated videos
2) AI generated videos
3) Live streams

1) Human generated videos: Video forensics field has been using Gen-AI tools to extract information from videos to detect threats. This includes processing individual frames as images and detecting objects and actions in those frames, converting audio to text and processing that text with Gen-AI techniques for further analysis.

- Detecting violence
- Detecting hate symbols
- Detecting explicit content

Bi-Long Short Term Memory (Bi-LSTM) machine learning model combined with Convolutional Neural Networks (CNNs) are often used to detecting violence and other harmful behaviour in video content. Sentiment analysis of the audio in the video too can be used to detect harmful videos. While a lot of such technology is proprietary, the research community has built example datasets that allows us to train and test models.

The VSD benchmark is a collection of ground-truth files based on the extraction of violent events in movies and web videos, together with high-level audio and video concepts [24]. Real life violent situations dataset is another major dataset that provides real world violence videos [93]. Violence Detection, A Serious-Gaming Approach is a IEEE Dataport dataset uses a serious gaming approach to collect data on violent and non-violent actions [14].

2) AI generated videos: Gen-AI breakthroughs now allow people to generate videos simply by using text prompts. Models like Sora have achieved remarkable results. However, this unrestricted ability to create videos introduces user safety concerns. These generated videos present two main challenges.

- Detecting generate videos that claim to be real (deepfakes)
- Detecting harm in generated videos

Deepfakes are pretty common issue on internet for a long time. But Gen-AI tools have made it easier to generate even more realistic looking content. Several techniques are under research to prevent models from generating deepfakes and detection of deepfakes [86] [77] [22].

Detecting harm in generated videos is another important research problem. Since generated videos are not bound by the physics of real world, generated videos can cleverly twist certain elements of the video to evade detection of standard algorithms [57].

Another case is where real videos are modified seamlessly to add or remove elements using Gen-AI. This is a harder problem to solve with limited amount of data availability [81] [100] [57].

3) Live streaming: Live video feeds is a popular form of video content. Besides social media live streams, security camera footage, traffic camera feeds, live sports and gaming etc. are important sources of live stream videos [64].

The user safety challenge between live stream videos and other type of videos is that inference time available to detect harm in live streams is much lower. [64] provides an extensive survey of technologies that can be used for detecting harmful multi-modal content in live streams. Gupta at el [41] discussed quantum machine learning as another potential approach for processing live streams.

D. Audio

Traditional machine learning techniques are limited in their ability to handle fabricated and harmful audio content. This inability usually stems as traditional methods rely on the extraction of acoustic features in the spectral domain. But with the advent of latest tools for generating natural sounding voice, traditional methods are bound to fail.

Gen-AI techniques are able to overcome this by helping in dataset generation for advancing research. Datasets such as FakeAVCeleb [6], [55] and Joint Audio-Visual Deepfake [122] improve deepfake detection research by including video deepfakes with corresponding synthesized, lipsynced audio tracks or by integrating audio alteration. Further efforts, such as WaveFake [32], which contains 100K+ generated audio samples contribute substantially to building resources capable of addressing the problem of detecting deepfakes. PolyGlot-Fake [44], a multimodal and multilingual deepfake dataset covers 7 languages, and MLAAD (Multi-Language Audio Anti-Spoofing Dataset) expands audio spoofing dataset to 23 languages [70].

Gen-AI techniques are also able to detect hate speech in verbal data. In particular, conformer model, an architecture combining convolutional networks and transformers, improves automatic speech recognition with a word error rate (WER) of less than 2% [38] . Further, generative models are helping to fill the lack of audio only datasets. For example, An et al. used text to speech models to generate an audio only dataset, and a BERT based model for explainable hate speech detection directly from audio files [7].

E. Code

Gen-AI do not simply generate code or superficially understand its context. They clearly demonstrate the ability to process it, identify the relevant parts, and operate on them. This capability is demonstrated in detecting malicious code, that is deliberately obfuscated to prevent this from happening. This makes them effective in deobfuscating malicious scripts [79], cyber threat detection [29], and even understanding minified code [36].

Chuanbo et al. [47] systematically leverage ChatGPT-4 to process multimodal app data (i.e., textual descriptions and screenshots) to determine maturity rating of an app to keep children safe from age inappropriate apps.

IV. Section: Adversarial Gen-AI - How Gen-AI is impacts threat landscape

In this section we analyze how Gen-AI technologies are impacting the adversaries of user safety [92] [91]. Earlier sections detail the well-understood threats to user safety. However, to better detect and deter these threats, we must better understand how bad actors might use Gen-AI techniques in adversarial settings [9], [30], [35], [69].

Table I describes the current landscape of user safety threats, potential victims, harm done and the bottleneck that the bad actors face. In the subsequent sections we describe how Gen-AI impacts these dimensions.

A. Safety Violations at Scale

Online fraud poses a major threat to user safety, with bad actors frequently aiming for large-scale disruption. Technology facilitates these attacks by enabling the mass distribution of emails and text messages. However, a limiting factor for these bad actors lies in their limited resources to conduct intelligent conversations with each victim. Gen-AI allows bad actors to overcome this crucial limitation. By using Gen-AI, they can reduce human involvement in large-scale operations, decreasing the time and cost of their attacks.

Gen-AI techniques now empower bad actors to craft more convincing communications and tailor them to each user, making detection by spam and threat detection algorithms more difficult [23]. Additionally, Gen-AI enables large-scale content generation [106], facilitating the creation of fake news websites and blogs that produce convincing, yet deceptive, content [23]. Further, foreign language content creation (i.e. geo-targeting) which was always a natural barrier for bad actors can be easily overcome using Gen-AI technology [68].

Thus Gen-AI has an impact on bottlenecks described in I. Gen-AI might allow bad actors to operate at higher scale with higher quality of deceptive content and might also create new distribution channels of communication such as chatbots, content websites and generated videos.

B. Safety Violations with Feedback

Reinforced learning can also help bad actors create better user-violation models that help them create more targeted and effective campaigns. Bad actors also have data of their last successful and un-successful attacks which is not accessible to the security researchers. This data can be very valuable for such re-inforced learning with human feedback.

Target of attack	Individual	Society/Groups	Institutions
Type of Attack	Personal information theft or blackmail	Information manipulation/ Misinformation	Synthetic reality
Benefit to bad actor	Financial	Political benefits/ Financial benefit	Financial benefits / Geopolitical benefits
Impact on victim	Financial and reputation loss	Generally distributed and small loss of many victims	Financial loss, Loss of control
Bad actor bottleneck	Ability to scale bespoke communication of communication.	Distribution channels of misinformation	Distribution channel of misinformation
Mitigation Responsibility	Individual, Personal Devices, Companies responsible for data.	Media companies, Regulators, Social Media companies.	Governments and Corporations
Example:	Email scam leading to credit card theft.	Crypto pump and dump schemes, Memestocks	Spreading rumors that cause harm.

TABLE I
THREATS TO USER SAFETY

Captcha is a common technique used by social media and other applications to prevent automated systems from using a system. However, modern Gen-AI techniques have enabled bad actors to develop sophisticated captcha solvers [112].

C. Safety Violations with Personalized Content

Gen-AI techniques have improved bad actor's ability to generate highly personalized content. For example, Gen-AI techniques with a large content window can technically look at a victim's available information and create personalized email campaigns, websites, calls or even video messages for phishing or financial fraud based scams. Further Gen-AI allows bad actors to turn a normal phishing attack into a spear phishing attack where instead of targeting a large group of people with similar messages, a highly targeted strategy is crafted to target specific individual specific individual.

Gen-AI techniques can also be used for pretexting [10] based scams, in this for of attack an bad actor designs an attack plan for their target. They create a story around the facts they know about the individual. The end goal of this story is eventually to scam the victim using prior information. This is a complex form of personalized scam and traditionally requires lot of resources from bad actor. However, with Gen-AI such level of personalization is relatively simpler.

D. Second order attacks

Gen-AI techniques have opened up a new front of attacks. For example, in these attacks human behavior is manipulated to achieve a certain end result by massive fake news campaigns and/or social media activity. We coin this phenomenon as "synthetic reality" [105]. Synthetic reality can be seen as a second order effect of a misinformation or targeted campaign by bad actors. In such attacks it is not clear who is the victim; but as long as some people fall prey to such activity the bad actors benefit. Attacks like these are extremely well coordinated and require complex infrastructure. It involves creating fake news websites, social media bots and even human users to amplify a certain type of message to make it sound very real.

Another form of second order effect attacks is where bad actors develop a long term plan to trick the world's latest Gen-AI technologies. Bad actor develop a long term plan to create synthetic reality on the internet through various specially setup websites, synthetic social media posts and similar mechanisms. Through this approach they poison the data on which these techniques are built.

E. Gen-AI Violations

AI safety is an evolving field where in models have inbuilt mechanism to ensure that they are not being used for user-harm [121] [21]. However, when such inbuilt safety mechanisms in the model are broken, it is referred to as jail-breaking the model [95] [49]. It is an active field of research as well as an area where regulations are being sought [63] [89]. Mujumdar et. al [66] present a much detailed survey of how malicious actors might be gaining strategic advantage with the help of Gen-AI despite the AI safety barriers on these models.

V. SECTION: FUTURE PROSPECTS OF GEN-AI

Researchers are rapidly advancing Gen-AI techniques, actively exploring many areas. We expect to see exciting developments that improve user safety, and we will describe three of these areas in the following subsections.

A. Content understanding

Content understanding is key to ensuring user safety. Traditionally, human moderators, supported by machine learning technologies, have analyzed content to ensure user safety. This analysis often uses some form of Reinforcement Learning through Human Feedback (RLHF). This has been a challenge due to limited availability of training data [31].

Modern Gen-AI techniques offer complex models that inherently identify negative content. Institutions leveraging Gen-AI for user safety features no longer need to develop their own content understanding models. This can be done by off-the-shelf models with some fine tuning or prompt engineering [65]. Using Gen-AI also benefits the minorities and outliers as they are much better represented in the training data of Gen-AI as compared to the internal data of an institution [31] [45]. Chaudhary et al. [16] gives an example of how a carefully crafted prompts and an off-the-shelf Gen-AI API's alone can produce excellent content moderation service. The availability of general-purpose models allows for the deployment of much more sophisticated Gen-AI technologies at a lower cost and without specialized content understanding. [83]

B. Foundation Ensembles

One of the biggest disruptions is the emergence of sufficiently large and complex models that can handle many business use cases, replacing multiple smaller, specialized models.A single model can now detect phishing, spam, moderate content, detect fraud, and analyze sentiment. This simplification streamlines the training, fine-tuning, and deployment of these models, accelerating development and deployment processes. Consequently, system complexity and associated costs decrease significantly. [34] [15].

1) Harnessing multi-modality for better user safety: Section III discussed the user safety challenges across data modalities and how Gen-AI impacts them. Future user-safety research using Gen-AI should focus on harnessing the multimodality of data. This research should analyze text, images, audio, and video data holistically for better user safety, rather than treating them as independent inputs. [78]

Yang et al. [110] showed that auto-insurance fraud detection is improved with the help of multi-modal models rather than analyzing only structural data. In future, Goyal et al. [37] hypothesized that there would be more applications of multimodal AI in improving user safety. Further, Akhare et al. [5] provided a survey of machine learning techniques in area of user safety and show that multi-modal multi-objective evolutionary algorithms are more scalable and more precise than other traditional methods to analyze content to enhance user safety.

C. Gen-AI and second order harm prevention

Future user safety research will likely focus on preventing the ”second-order effects” described in section IV-D. Research like [54] have shown that long term psychological and financial harm is possible due to widespread Gen-AI usage. For example, Dewite [26] shows that extensive usage of chatbots might be harmful in long term even though independently specific conversations with the chatbot might appear harmless. Markus et al. [8] shows that how certain AI uses might be perfectly normal independently but when chained together can create a misuse chain and how this might be detected.

User safety researchers might have to take more holistic approach towards user safety instead of restricting themselves with specific methods or modalities [18]. Kranz et al. [61] demonstrated how AI-copilot can be provided to humans in their interactions with robots to ensure safety, by flagging unusual behaviour.

We anticipate the widespread adoption of co-pilots that monitor users' online behavior across platforms. These copilots will provide a comprehensive safety mechanism, protecting users from a wide variety of second-order effects stemming from the pervasive use of Gen-AI in everyday life. Such technologies will detect synthetic realities and prevent user from fall falling prey to complex second order effect harms.

VI. Section: Conclusion

Gen-AI techniques, along with there inherent ability w.r.t translation between languages, fine-tuning between various tasks and domains [25] are able to overcome the challenges associated with ML/DM techniques to reduce violation w.r.t user-safety tasks.

In this survey, we provided an overview of the various user safety domains (e.g. egregious abuse, misinformation and disinformation, increase in content moderation, towards robust physical safety) across which user safety can be violated and how Gen-AI techniques can be used to overcome those avenues. Next, we discussed how Gen-AI techniques can be used across various data modalities i.e. text, audio, video, executable binaries and images.

We then discussed how Gen-AI techniques can be used in an adversarial setting. In particular, we discussed how these techniques can be used to pursue safety violations at scale, safety violations with human feedback, safety violations with personalized content, second order attacks and Gen-AI violations.

Lastly, we provide our opinion about the future prospects of Gen-AI techniques w.r.t user-safety. In particular, we hypothesize how Gen-AI techniques have an inherent ability w.r.t. content understanding, their ability to work as large ensembles with one model pursuing multiple tasks and lastly how Gen-AI techniques can prevent second order harm.

We believe that this work represents the first summarization of Gen-AI techniques for user-safety. In particular, we provided a deep overview of emerging technologies along with their applications to user-safety, with a focus on areas suitable for advancement. Our goal is to make sure that this work warrants immediate investment towards driving growth within the industry.

References

[1] Swati Agrawal and Ashish Sureka. Copyright infringement detection of music videos on youtube by mining video and uploader meta-data. In *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings 2*, pages 48–67. Springer, 2013.

[2] Syed Hammad Ahmed, Shengnan Hu, and Gita Sukthankar. The potential of vision-language models for content moderation of children's videos, 2023.

[3] Syed Hammad Ahmed, Muhammad Junaid Khan, Hafiz Muhammad Umer Qaisar, and Gita Reese Sukthankar. Malicious or benign? towards effective content moderation for children's videos. *ArXiv*, abs/2305.15551, 2023.

[4] Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. Defending against social engineering attacks in the age of llms, 2024.

[5] Vishakha D Akhare and LK Vishwamitra. Machine learning models for fraud detection: A comprehensive review and empirical analysis. *Journal of Electrical Systems*, 20(3s):1138–1149, 2024.

[6] Abdulazeez Alali and George Theodorakopoulos. Review of existing methods for generating and detecting fake and partially fake audio. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, pages 35–36, 2024.

[7] Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. An investigation into explainable audio hate speech detection. *arXiv preprint arXiv:2408.06065*, 2024.

[8] Markus Anderljung and Julian Hazell. Protecting society from ai misuse: when are restrictions on capabilities warranted? *arXiv preprint arXiv:2303.09377*, 2023.

[9] Meraj Farheen Ansari, Pawan Kumar Sharma, and Bibhu Dash. Prevention of phishing attacks using ai-based cybersecurity awareness training. *Prevention*, 3(6):61–72, 2022.

[10] Subia Ansari. From the scammer perspective: Predispositions towards online fraud motivation and rationalization. Master's thesis, Purdue University, 2020.

[11] Ahmed Bensaoud, Jugal Kalita, and Mahmoud Bensaoud. A survey of malware detection using deep learning. *Machine Learning with Applications*, 16:100546, June 2024.

[12] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[13] Le Binh, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, Simon S. Woo, and Jelena Mirkovic. Samba: Identifying inappropriate videos for young children on youtube. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.

[14] giuseppe cascavilla and Stan Ruessink. Violence detection: A serious-gaming approach, 2023.

[15] Joymallya Chakraborty, Wei Xia, Anirban Majumder, Dan Ma, Walid Chaabene, and Naveed Janvekar. Detoxbench: Benchmarking large language models for multitask fraud & abuse detection. *arXiv preprint arXiv:2409.06072*, 2024.

[16] Tarun Kumar Chawdhury. Content moderation: An llm api with a carefully crafted system prompt is all you need. 2024.

[17] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7621–7639. PMLR, 21–27 Jul 2024.

[18] Chen Chen, Ziyao Liu, Weifeng Jiang, Goh Si Qi, and KwoK-Yan Lam. Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*, 2024.

[19] Zhiling Chen, Hanning Chen, Mohsen Imani, Ruimin Chen, and Farhad Imani. Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces. *arXiv preprint arXiv:2408.07146*, 2024.

[20] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study. *arXiv preprint arXiv:0707.3670*, 2007.

[21] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*, 2024.

[22] Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. As good as a coin toss human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760*, 2024.

[23] Audrey de Rancourt-Raymond and Nadia Smaili. The unethical use of deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

[24] Claire-Hélène Demarty, Cédric Penet, Mohammad Soleymani, and Guillaume Gravier. Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. *Multimedia Tools and Applications*, 74:7379–7404, 2015.

[25] Akshar Prabhu Desai, Ganesh Satish Mallya, Mohammad Luqman, Tejasvi Ravi, Nithya Kota, and Pranjul Yadav. Opportunities and challenges of generative-ai in finance, 2024.

[26] Pierre Dewitte. Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review*, 54:106019, 2024.

[27] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[28] Mohamed Amine Ferrag, Fatima Alwahedi, Ammar Battah, Bilel Cherif, Abdechakour Mechri, and Norbert Tihanyi. Generative ai and large language models for cyber security: All insights you need, 2024.

[29] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C. Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *IEEE Access*, 12:23733–23750, 2024.

[30] Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, pages 1–21, 2024.

[31] Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 1–8, 2023.

[32] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

[33] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection, 2021.

[34] Sean Gallagher, Ben Gelman, Salma Taoufiq, Tamás Vörös, Younghoo Lee, Adarsh Kyadige, and Sean Bergeron. Phishing and social engineering in the age of llms. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 81–86. Springer Nature Switzerland Cham, 2024.

[35] Kotie Geldenhuys. The darker side of artificial intelligence. *Servamus Community-based Safety and Security Magazine*, 116(11):20–25, 2023.

[36] Glama. Openai is shockingly good at unminifying code. https://glama.ai/blog/2024-08-29-reverse-engineering-minified-code-using-openai.

[37] Bharti Goyal, Nasib Singh Gill, Preeti Gulia, Om Prakash, Ishaani Priyadarshini, Rohit Sharma, Ahmed J Obaid, and Kusum Yadav. Detection of fake accounts on social media using multimodal data with deep learning. *IEEE Transactions on Computational Social Systems*, 2023.

[38] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[39] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE, 2023.

[40] Keyan Guo, Ayush Utkarsh, Wenbo Ding, Isabelle Ondracek, Ziming Zhao, Guo Freeman, Nishant Vishwamitra, and Hongxin Hu. Moderating illicit online image promotion for unsafe user-generated content games using large vision-language models. *arXiv preprint arXiv:2403.18957*, 2024.

[41] Geetika Gupta, Karuna Kadian, Raksha Jain, Vimal Dwivedi, and Arun Sharma. Real-time hate speech detection in live streaming platforms using quantum machine learning. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE, 2023.

[42] Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. Preserving integrity in online social networks, 2020.

[43] Seth Hardy, Masashi Crete-Nishihata, Katharine Kleemola, Adam Senft, Byron Sonne, Greg Wiseman, Phillipa Gill, and Ronald J. Deibert. Targeted threat index: characterizing and quantifying politically-motivated targeted malware. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, page 527–541, USA, 2014. USENIX Association.

[44] Yang Hou, Haitao Fu, Chuankai Chen, Zida Li, Haoyu Zhang, and Jianjun Zhao. Polyglotfake: A novel multilingual and multimodal deepfake dataset. *arXiv preprint arXiv:2405.08838*, 2024.

[45] Arkar Htet, Sui Reng Liana, Theingi Aung, and Amiya Bhaumik. Chatgpt in content creation: Techniques, applications, and ethical implications. In *Advanced Applications of Generative AI and Natural Language Processing Models*, pages 43–68. IGI Global, 2024.

[46] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113, March 2024.

[47] Chuanbo Hu, Bin Liu, Minglei Yin, Yilu Zhou, and Xin Li. Multimodal chain-of-thought reasoning via chatgpt to protect children from age-inappropriate apps. *arXiv preprint arXiv:2407.06309*, 2024.

[48] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual

program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024.
[49] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.
[50] Hochul Hwang, Sunjae Kwon, Yekyung Kim, and Donghyun Kim. Is it safe to cross? interpretable risk assessment with gpt-4v for safety-aware street crossing. *arXiv preprint arXiv:2402.06794*, 2024.
[51] Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Jhuma kabir Mim, and Nabil Arhab. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms, 2024.
[52] Christian Jansohn, Adrian Ulges, and Thomas M Breuel. Detecting pornographic video content by combining image features with motion information. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 601–604, 2009.
[53] Rajeshwari Kandakatla. Identifying offensive videos on youtube. Master's thesis, Wright State University, 2016.
[54] Ferial Khaddage and Walid Safi. The good the bad the ugly of artificial intelligence & why it matters in education. In *EdMedia+ Innovate Learning*, pages 1791–1796. Association for the Advancement of Computing in Education (AACE), 2019.
[55] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.
[56] Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. Llms in the loop: leveraging large language model annotations for active learning in low-resource languages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 397–412. Springer, 2024.
[57] Staffy Kingra, Naveen Aggarwal, and Nirmal Kaur. Emergence of deepfakes and video tampering detection approaches: A survey. *Multimedia Tools and Applications*, 82(7):10165–10209, 2023.
[58] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Detecting phishing sites using chatgpt. *ArXiv*, abs/2306.05816, 2023.
[59] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection, 2024.
[60] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.
[61] Philipp Kranz, Fabian Schirmer, Tobias Kaupp, and Marian Daun. Generative ai co-pilot to support safety analyses of human-robot collaborations. *IEEE Software*, 2024.
[62] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
[63] Yuxi Li, Yi Liu, Yuekang Li, Ling Shi, Gelei Deng, Shengquan Chen, and Kailong Wang. Lockpicking llms: A logit-based jailbreak using token-level manipulation. *arXiv preprint arXiv:2405.13068*, 2024.
[64] Helena Liz-Lopez, Mamadou Keita, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Javier Huertas-Tato, and David Camacho. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, 103:102103, 2024.
[65] Yaaseen Mahomed, Charlie M Crawford, Sanjana Gautam, Sorelle A Friedler, and Danaë Metaxa. Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 660–686, 2024.
[66] Durjoy Majumdar, S Arjun, Pranavi Boyina, Sri Sai Priya Rayidi, Yerra Rahul Sai, and Suryakanth V Gangashetty. Beyond text: Nefarious actors harnessing llms for strategic advantage. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, pages 1–7. IEEE, 2024.
[67] Samuel Marchal, Kalle Saari, Nidhi Singh, and N. Asokan. Know your phish: Novel techniques for detecting phishing sites and their targets, 2016.
[68] Alakananda Mitra, Saraju P Mohanty, and Elias Kougianos. The world of generative ai: Deepfakes and large language models. *arXiv preprint arXiv:2402.04373*, 2024.
[69] Priyabrata Mohapatra, Diddigi Kulkarni Nitish, and Isah Kaura Shehu. Minds of malice: Unraveling the web of ai crime. *The Scientific Spectrum of AI Enhancing The Future*, page 155.
[70] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv:2401.09512*, 2024.
[71] N Nagashree, Ravi Tejasvi, and KC Swathi. An early risk detection and management system for the cloud with log parser. *Computers in Industry*, 97:24–33, 2018.
[72] CBS News. Ai counterfeit detection: Amazon's new tool in the fight against fake products, Oct 2024.
[73] Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts, 2023.
[74] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.
[75] Hakan T. Otal and M. Abdullah Canbaz. Ai-powered crisis response: Streamlining emergency management with llms. In *2024 IEEE World Forum on Public Safety Technology (WFPST)*, pages 104–107, 2024.
[76] Hakan T. Otal, Eric Stern, and M. Abdullah Canbaz. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859, 2024.
[77] Yan Pang, Aiping Xiong, Yang Zhang, and Tianhao Wang. Towards understanding unsafe video generation. *arXiv preprint arXiv:2407.12581*, 2024.
[78] Jongchan Park, Min-Hyun Kim, and Dong-Geol Choi. Correspondence learning for deep multi-modal recognition and fraud detection. *Electronics*, 10(7):800, 2021.
[79] Constantinos Patsakis, Fran Casino, and Nikolaos Lykousas. Assessing llms in malicious code deobfuscation of real-world malware campaigns. *Expert Systems with Applications*, 256:124912, 2024.
[80] Jianbiao Peng, Beiji Zou, and Chengzhang Zhu. Combining external attention gan with deep convolutional neural networks for real–fake identification of luxury handbags. *The Visual Computer*, 39(3):971–982, 2023.
[81] Armaan Pishori, Brittany Rollins, Nicolas van Houten, Nisha Chatwani, and Omar Uraimov. Detecting deepfake videos: An analysis of three techniques. *arXiv preprint arXiv:2007.08517*, 2020.
[82] Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320, 2020.
[83] Wei Qiao, Tushar Dogra, Otilia Stretcu, Yu-Han Lyu, Tiantian Fang, Dongjin Kwon, Chun-Ta Lu, Enming Luo, Yuan Wang, Chih-Chun Chia, et al. Scaling up llm reviews for google ads content moderation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1174–1175, 2024.
[84] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486*, 2024.
[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
[86] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.
[87] Jonas Rieskamp, Milad Mirbabaie, and Kerstin Zander. Genai-powered social bots for crisis communication: A systematic literature review. 2023.
[88] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[89] Felipe Romero Moreno. Generative ai and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, pages 1–30, 2024.

[90] Siva Sai, Alfred W Jacob, Sakshi Kalra, and Yashvardhan Sharma. Stacked embeddings and multiple fine-tuned xlm-roberta models for enhanced hostility identification. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 224–235. Springer, 2021.

[91] Wissam Salhab, Darine Ameyed, Fehmi Jaafar, and Hamid Mcheick. A systematic literature review on ai safety: Identifying trends, challenges and future directions. *IEEE Access*, 2024.

[92] Marc Schmitt and Ivan Flechais. Digital deception: Generative artificial intelligence in social engineering and phishing. *arXiv preprint arXiv:2310.13715*, 2023.

[93] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *2019 ninth international conference on intelligent computing and information systems (ICICIS)*, pages 80–85. IEEE, 2019.

[94] Micol Spitale, Minja Axelsson, and Hatice Gunes. Vita: A multi-modal llm-based system for longitudinal, autonomous, and adaptive robotic mental well-being coaching, 2023.

[95] Jingtong Su, Julia Kempe, and Karen Ullrich. Mission impossible: A statistical perspective on jailbreaking llms. *arXiv preprint arXiv:2408.01420*, 2024.

[96] Jinyan Su, Claire Cardie, and Preslav Nakov. Adapting fake news detection to the era of large language models, 2024.

[97] Wei-Lun Tsai, Jacob J Lin, Wang-Fat Ho, Shuai Tang, and Shang-Hsien Hsieh. Construction safety inspection workflow with clip-based image captioning and attention generation. *Available at SSRN 4819831*.

[98] Pichapa Vanpech, Kanchicha Peerabenjakul, Napatsawan Suriwong, and Somchart Fugkeaw. Detecting cyberbullying on social networks using language learning model. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*, pages 161–166, 2024.

[99] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*, 2021.

[100] Ainuddin Wahid Abdul Wahab, Mustapha Aminu Bagiwa, Mohd Yamani Idna Idris, Suleman Khan, Zaidi Razak, and Muhammad Rezal Kamel Ariffin. Passive video forgery detection techniques: A survey. In *2014 10th International conference on information assurance and security*, pages 29–34. IEEE, 2014.

[101] Han Wang, Qiaozhi Bao, Zuwei Shui, Lianwei Li, and Huan Ji. A novel approach to credit card security with generative adversarial networks and security assessment. 2024.

[102] Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation, 2024.

[103] Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, and Chuang Zhang. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation. *ArXiv*, abs/2403.14171, 2024.

[104] Tongze Wang, Xiaohui Xie, Lei Zhang, Chuyi Wang, Liang Zhang, and Yong Cui. Shieldgpt: An llm-based framework for ddos mitigation. In *Proceedings of the 8th Asia-Pacific Workshop on Networking*, APNet '24, page 108–114, New York, NY, USA, 2024. Association for Computing Machinery.

[105] Y. Wang. Synthetic realities in the digital age: navigating the opportunities and challenges of ai-generated content. 2023.

[106] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378, 2024.

[107] Siyi Wu, Julie Y. A. Cachia, Feixue Han, Bingsheng Yao, Tianyi Xie, Xuan Zhao, and Dakuo Wang. "i like sunnie more than i expected!": Exploring user expectation and perception of an anthropomorphic llm-based conversational agent for well-being support, 2024.

[108] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation, 2024.

[109] Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulcar, Basant Agarwal, and Marko Robnik Sikonja. Code-mixed sentiment and hate-speech prediction. *arXiv preprint arXiv:2405.12929*, 2024.

[110] Jiaxi Yang, Kui Chen, Kai Ding, Chongning Na, and Meng Wang. Auto insurance fraud detection with multimodal learning. *Data Intelligence*, 5(2):388–412, 2023.

[111] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[112] Guixin Ye, Zhanyong Tang, Dingyi Fang, Zhanxing Zhu, Yansong Feng, Pengfei Xu, Xiaojiang Chen, Jungong Han, and Zheng Wang. Using generative adversarial networks to break and protect text captchas. *ACM Transactions on Privacy and Security (TOPS)*, 23(2):1–29, 2020.

[113] Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. Evidence-driven retrieval augmented response generation for online misinformation. In *North American Chapter of the Association for Computational Linguistics*, 2024.

[114] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. Detecting deepfake videos with temporal dropout 3dcnn. In *International Joint Conference on Artificial Intelligence*, 2021.

[115] Daniel Yue Zhang, Jose Badilla, Herman Tong, and Dong Wang. An end-to-end scalable copyright detection system for online video sharing platforms. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 626–629. IEEE, 2018.

[116] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[117] Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *International Joint Conference on Natural Language Processing*, 2023.

[118] Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*, 2024.

[119] Haonan Zhao, Yiting Wang, Thomas Bashford-Rogers, Valentina Donzella, and Kurt Debattista. Exploring generative ai for sim2real in driving data synthesis. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, volume 34, page 3071–3077. IEEE, June 2024.

[120] Wenxiang Zhao, Juntao Wu, and Zhaoyi Meng. Apppoet: Large language model based android malware detection via multi-view prompt engineering, 2024.

[121] Weikang Zhou, Xiao Wang, Limao Xiong, Han Xia, Yingshuang Gu, Mingxu Chai, Fukang Zhu, Caishuang Huang, Shihan Dou, Zhiheng Xi, et al. Easyjailbreak: A unified framework for jailbreaking large language models. *arXiv preprint arXiv:2403.12171*, 2024.

[122] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.