CERT-EU

Generative AI in Cybersecurity

Balancing Innovation and Risk

CERT-EU Team
ver. **2.0**
July 3, 2025

Contents

1 Introduction

Some portions of this document were crafted with a little help from our internally deployed AI models (powered by Llama 3.3 - 70B). While we have fine-tuned and polished the content, this also showcases how Generative AI can be a powerful tool – especially when running on your own infrastructure for greater control, security and privacy.

1.1 Generative AI

Generative AI refers to a class of artificial intelligence models designed to simulate human-like creativity and adaptability by generating new content, data, or outputs based on learned patterns from vast datasets. These AI systems can be applied across a wide range of domains, from natural language processing to computer vision. For example, Generative AI can create realistic images, draft human-like text, compose music, or design novel chemical compounds. Notable examples include large language models (LLMs) such as OpenAI's GPT series, Mistral Le Chat, Google's Gemini, and Meta's LLaMA, as well as text-to-image generation tools like OpenAI's DALL-E and Stability AI's Stable Diffusion.

Large language models (LLMs) are advanced deep learning models that use self-attention mechanisms and multi-layered architectures to understand and generate text. These models excel in tasks such as language translation, summarisation, question-answering, and creative content generation by analysing vast quantities of data and identifying complex patterns. Their strengths lie in their ability to capture nuanced contextual information, generate coherent and relevant responses, and adapt to diverse domains.

However, LLMs also have notable limitations. They are highly data- and computationally intensive, requiring substantial resources for training and fine-tuning. Additionally, they may produce plausible-sounding yet incorrect or nonsensical answers (often referred to as hallucinations) and can be sensitive to input phrasing, leading to inconsistent results. Finally, they may inadvertently generate biased or harmful content due to biases present in their training data.

Similarly, text-to-image generation models utilise deep learning techniques to create visually coherent images based on textual input. The strengths of these models include their ability to generate diverse and creative images, as well as contributing to data augmentation and visual storytelling. However, like LLMs, their weaknesses include a dependence on large, well-annotated datasets for training, high computational requirements, and the potential to generate unrealistic or low-quality images. Furthermore, these models may struggle to accurately capture complex and abstract concepts described in the textual input and, similar to transformers, may inadvertently propagate biases present in the training data.

1.2 How does it work?

Generative AI might seem like magic at first, but it's actually the result of significant progress in deep learning. This progress is driven by rapid increases in computing power, access to large datasets, and better training techniques such as reinforcement learning and self-supervised learning. These advances have come together to make Generative AI a reality, rather than just an idea.

1.2.1 Text generation models

At the core of large language models are neural networks with millions or even billions of parameters (i.e., *model size*), which are trained on vast amounts of text data. These *parameters*

define the connections between the nodes in the network. Before an LLM can be used, it must undergo a training process, during which it is presented with massive datasets (known as training sets) that allow the model to learn patterns, relationships, and structures within the language. Through this, the model adjusts its parameters to minimise errors in prediction. Modern LLMs are trained using unsupervised learning and self-supervised learning, where the model learns by predicting the next word (or more precisely: *token*) in a sequence based on the prior context. LLMs can be broadly categorised into:

- **Foundational models**

  Foundational models, such as OpenAI's GPT series, Anthropic's Claude, Meta's LLaMA, and others – serve as the base for a wide range of applications. These models are pre-trained on enormous, diverse datasets sourced from books, articles, websites, and other publicly available texts. Rather than learning specific tasks, foundational models learn a probabilistic distribution of language – in other words, they grasp how words, phrases, and concepts typically relate to one another across a wide array of contexts. They are general-purpose models capable of understanding and generating text across numerous domains without being tailored to specific tasks. Foundational models are highly versatile and can be applied to various tasks, including translation, summarisation, creative writing, and even code generation.

- **Fine-tuned models**

  Fine-tuned models build upon foundational models but are adapted for specific tasks or domains. They undergo additional training, or fine-tuning, using task-specific data or techniques such as Reinforcement Learning from Human Feedback (RLHF). Fine-tuning enhances the model's performance in particular applications, allowing it to better meet user needs. For instance, OpenAI's ChatGPT (based on the GPT-4 model) is a fine-tuned version of a foundational model, optimised for conversational AI. It has been trained not only to generate text but also to handle dialogue, ensuring that responses are contextually relevant, informative, and aligned with user intent. This fine-tuning process helps reduce errors and improves the model's reliability in real-world applications, such as customer service, virtual assistants, and personalised content generation.

Once training is complete, the model enters the *inference* phase, where it generates predictions or completes tasks based on new input data. During this stage, the model leverages its internal knowledge of language patterns and relationships, acquired during training, to produce relevant and coherent output. For foundational models, this means generating responses based on a broad understanding of language and context, while fine-tuned models use additional task-specific training to generate more tailored and accurate responses.

1.2.2 Visual generation models

Generative AI models are used to create visual content, such as images and videos, from textual descriptions. These models use neural networks to produce high-quality visual outputs. They are changing the creative industries by allowing dynamic content to be generated, whether it's a single image or a video sequence.

There are different types of models used for image and video generation. For example, Generative Adversarial Networks (GANs) and Diffusion Models, like DALL-E 3 and Stable Diffusion, are commonly used. Video models are more complex because they need to maintain both spatial and temporal coherence.

Image generation models focus on converting text prompts into static images. Diffusion Models have become popular in recent years and work by refining random noise into structured images

using text prompts. To train these models, large datasets of paired text and image data are needed. This allows the model to learn relationships between textual descriptions and their corresponding visual representations.

Video generation builds on the foundations of image generation but introduces the challenge of spatio-temporal relationships. This means ensuring that individual frames and transitions between them are coherent over time. Video generation models often adapt Diffusion Models to handle sequential frame generation while considering temporal flow. They need to learn to generate high-quality images for each frame and ensure movements, lighting, and objects remain consistent throughout the sequence.

Both image and video generation models rely on the concept of latent space during inference. Latent space is an abstract representation of the learned relationships between visual and textual elements. In image generation, the model samples from this latent space to produce a single visual output. For video generation, the latent space also encodes temporal dynamics, enabling the model to generate a sequence of frames that align with the input text while ensuring smooth transitions.

To achieve this, video generation models are trained on vast video-text paired datasets. The focus is on learning to generate not just realistic images but also seamless motion and narrative progression. This results in dynamic content that feels natural while being driven by the input text. By understanding how these models work, you can explore their potential applications in various industries and creative projects.

1.3 Future outlook

Artificial intelligence is advancing rapidly, with transformative innovations reshaping industries and unlocking new opportunities. The growth of open-source and open-weights models has significantly expanded AI's accessibility and applications. Unlike proprietary systems, these frameworks allow organisations, researchers, and developers to deploy, adapt, and refine AI tools without restrictive licensing, promoting greater autonomy. This shift is particularly evident in Generative AI, where advanced open source projects such as DeepSeek are driving significant breakthroughs. These models now enable diverse applications, from language processing to image and video generation, empowering businesses and individuals to innovate cost-effectively.

Cost-effective AI models are now rivaling proprietary models in performance. Research into self-supervised, unsupervised, and reinforcement learning is advancing rapidly, while breakthroughs in multimodal AI – integrating text, images, audio, video, and even interactive environments – are pushing the boundaries of creativity and problem-solving.

At CERT-EU, we are pursuing in-house AI projects to optimise operations and deepen our expertise in these technologies. This hands-on approach ensures that our advice on AI systems is both informed and actionable. While AI's democratisation offers opportunities to enhance innovation, personalise experiences, and automate tasks, it also presents significant risks, such as the spread of disinformation, misuse by malicious actors, and ethical dilemmas surrounding synthetic content. As the technology evolves, it is crucial to address these challenges proactively to ensure its responsible development.

2 The dual edge of Generative AI

Generative AI holds immense transformative potential, reshaping various sectors, including cybersecurity. This document specifically focuses on its implications for Union entities, examining how Generative AI is revolutionising both the defensive and offensive aspects of cybersecurity.

By exploring how these technologies can strengthen protective measures while also enabling new threats, we aim to assess their impact on organisations within the Union.

Our goal is to propose actionable recommendations that will help direct and coordinate the efforts of Union entities in effectively harnessing the benefits of Generative AI, while mitigating its associated risks. Given the rapid pace of innovation, these insights reflect the landscape as of mid-2025, with the understanding that ongoing developments may substantially alter the threat environment.

2.1 Defensive applications of AI

Artificial intelligence offers powerful tools to counter sophisticated cyber threats, mitigating the traditional "Defender's Dilemma"[1] – where attackers historically retain the upper hand. AI systems excel at transforming vast datasets into actionable intelligence, bolstering capabilities such as malware detection, vulnerability identification, and threat analysis. By automating routine tasks and accelerating response times, these technologies enable security teams to operate with heightened efficiency and precision.

Generative AI further augments defences. For example, it can simulate realistic cyber-attack scenarios, such as phishing campaigns or ransomware simulations, to rigorously test and train personnel. Beyond training, generative models facilitate the creation of adaptive honeypots that mislead attackers while gathering tactical intelligence. Large language models add value by detecting subtle patterns in data, such as log file anomalies, empowering analysts to prioritise risks and uncover hidden correlations.

2.2 Offensive applications of AI

The same technologies empowering defenders are increasingly weaponised by adversaries. Generative AI, for instance, is exploited to craft hyper-realistic social engineering campaigns, including personalised phishing emails, SMS scams, and deepfake audio/video. Beyond deception, these models automate the discovery of software vulnerabilities – even uncovering novel attack vectors, and generate functional code for malware or evasion techniques[2]. Tools like WormGPT or FraudGPT, for example, illustrate how readily AI can be adapted to scale malicious activities.

The escalation of AI-driven disinformation and cyberattacks presents acute challenges for the Union entities. This evolving threat landscape underscores the urgent need for proactive, AI-enhanced security strategies, such as behaviour-based anomaly detection and predictive threat modelling, to counter adversarial innovation.

It is essential for CERT-EU to communicate not only the risks posed by Generative AI but also the opportunities it offers to enhance the resilience and cybersecurity capabilities of Union entities. Collaboration among these entities will be crucial for sharing best practices, investing in development, and establishing ethical frameworks for AI use. By prioritising human-AI collaboration and maintaining rigorous oversight, CERT-EU and Union entities can set a benchmark, fostering a secure and innovative digital ecosystem within the public institutions of the European Union.

3 Adoption of AI in cybersecurity

As AI continues to evolve, organisations are now focused on integrating these technologies into their existing cybersecurity teams and systems. The challenge lies in ensuring the effective

[1] *The Defender's Dilemma*: "Defenders have to be right every time. Attackers only need to be right once."
[2] ENISA Threat Landscape 2024

adoption of AI solutions, aligning them with strategic objectives, and optimising overall security efforts.

A 2024 survey by the Cloud Security Alliance (CSA)[3] found that 63% of security professionals believe AI will enhance security measures. However, adoption remains in its early stages, with only 22% of organisations currently using Generative AI, though 55% plan to implement it within the next year.

3.1 AI's impact on security operations

AI-driven automation is reshaping security operations by improving threat detection, accelerating incident response, and reducing operational burdens. While concerns about AI replacing human roles exist, most professionals see it as a complementary tool. Only 12% believe AI will fully replace their roles, whereas 30% expect it to enhance their skills, and 28% see AI as broadly supporting their work. However, 51% warn against over-reliance, emphasising the need for human oversight in AI-driven security strategies.

Organisations are deploying AI to address workforce shortages and improve security efficiency. Around 36% of security teams use AI to bridge skill gaps, while 26% prioritise faster threat detection. Other key objectives include improving productivity and reducing misconfigurations. Generative AI is increasingly used for automated rule creation, attack simulations, and compliance monitoring, but its adoption introduces risks such as data manipulation and adversarial attacks. Effective governance frameworks are necessary to ensure responsible AI deployment.

3.2 Challenges in AI adoption

Despite its advantages, AI adoption in cybersecurity presents significant challenges. A primary concern is the shortage of skilled professionals who can effectively manage and secure AI systems, with 33% of organisations citing skill gaps as a major barrier. The complexity of AI models requires specialised expertise to train, maintain, and interpret their outputs. Additionally, 38% of security professionals highlight data quality issues, including unintended bias in AI models, while 25% cite privacy risks as a growing concern.

AI security risks extend beyond implementation challenges. Security professionals are increasingly aware of vulnerabilities such as data poisoning, adversarial attacks, and AI-generated misinformation. Approximately 28% of respondents express concerns about data poisoning attacks, which could manipulate AI models to generate false outputs. Regulatory and compliance considerations further complicate AI deployment, requiring organisations to navigate evolving security frameworks.

3.3 Future trends in AI-Driven cybersecurity

AI adoption is expected to accelerate, with organisations increasingly exploring Generative AI for cybersecurity applications. Automated rule creation, attack simulations, and compliance monitoring are among the most common use cases. While AI-driven automation will continue to expand, security teams must balance efficiency gains with oversight mechanisms to mitigate potential risks.

Governance structures are evolving to support responsible AI deployment. Many organisations are establishing dedicated teams to oversee AI implementations, ensuring compliance with security policies and regulatory requirements. Executive leadership plays a crucial role in AI adoption, with 82% of organisations reporting strong leadership support for AI initiatives. However,

[3] CSA State of the AI

a gap remains between executive-level strategy and operational execution, reinforcing the need for clear guidelines and AI-specific training programmes.

As AI adoption progresses, cybersecurity strategies must adapt to emerging threats. The increasing sophistication of AI-driven cyberattacks necessitates continuous advancements in AI-based defence mechanisms. Security teams must also address AI security risks, improve transparency in AI decision-making, and implement safeguards against adversarial manipulation. The widespread integration of AI into cybersecurity marks a transformative shift, requiring a strategic approach to maximise its benefits while mitigating associated risks.

3.4 European Union initiatives and investments in AI

The European Commission has made significant strides in promoting the secure and compliant use of Generative AI technologies. A notable example is its approval of the Azure OpenAI service for handling Sensitive Non-Classified Data, which ensures that user prompts remain within the EU and are excluded from model training. Hosted in regional cloud environments under Cloud Broker Contracts, this service strikes a balance by providing robust privacy safeguards without the complexities and resource-intensive demands of managing local models.

Similarly, the Commission has implemented GPT@EC[4], deploying AI models within a secure, local infrastructure to enhance operational efficiency while adhering to EU data protection standards. Complementing these efforts, CERT-EU has integrated internal large language models to power AI-driven projects, ensuring sensitive data remains protected while boosting productivity and supporting internal initiatives to reduce manual workloads and enhance decision-making processes

Looking ahead, the European Union has pledged €200 billion in AI investments[5], aiming to build critical infrastructure such as AI training gigafactories within the region. This ambitious plan highlights Europe's commitment to advancing AI innovation while maintaining a strong focus on data privacy and regulatory compliance.

4 Benefits of using Generative AI

As this document focuses on the impact of Generative AI on cybersecurity and its relevance to Union entities, several updated use cases in cybersecurity are highlighted, with an emphasis on ongoing AI-driven initiatives.

4.1 Improving threat detection

AI systems enhance cybersecurity by refining detection algorithms and generating new rules based on the latest threat intelligence. By leveraging AI-driven insights, security teams can improve detection capabilities, strengthening defences against cyberattacks.

Log analysis is a critical aspect of a cybersecurity analyst's role, and large language models (LLMs) can significantly streamline this process. AI can sift through vast amounts of log data, identifying anomalies, outliers, and correlations that may indicate security breaches. Automating log analysis reduces the time and effort required, allowing analysts to focus on high-priority threats rather than manually searching for patterns. Additionally, AI systems can detect connections between seemingly unrelated events, providing a comprehensive view of the security landscape and enabling more effective threat responses.

[4] GPT@EC
[5] EU launches InvestAI

AI also holds promise in enhancing detection rules by analysing large datasets, such as system logs, to uncover patterns that may be missed by human analysts. However, false positives remain a major concern. Even with a low false positive rate of 0.1%, the sheer volume of log entries – often in the hundreds of millions – can result in an overwhelming number of alerts, necessitating careful tuning of AI models.

CERT-EU is exploring the use of retrieval-augmented generation (RAG)-based systems to assist analysts in creating and refining detection rules. These systems leverage existing rules to generate new ones, which analysts can review and implement. Other AI-driven platforms already being used for threat detection and response include Microsoft Security Copilot[6] and Palo Alto Networks' Cortex XSOAR[7], demonstrating AI's growing role in strengthening cybersecurity operations.

4.2 Supporting analysis

LLMs have also enabled analysts to work more efficiently and accurately in various aspects of their jobs. One key area where they have demonstrated their potential is in the deobfuscation of malicious code. Attackers often obfuscate their code to evade detection, but LLMs can assist analysts in identifying patterns and decoding hidden algorithms, providing valuable insights into the attacker's intent and revealing the true nature of the threat, thereby significantly speeding up the investigation.

Additionally, LLM transformers show a remarkable ability to correlate data from various sources and fields as they have been trained on diverse datasets. This vast training corpus enables the AI model to extract and synthesise information from various seemingly unrelated sources. By leveraging its deep learning capabilities, LLMs can then identify connections, patterns, and insights across these different sources. As a result, the AI model has proven to be a valuable tool in solving problems, providing novel insights, and identifying correlations that would otherwise be easy to miss.

Code analysis and reverse engineering are also areas where LLMs can provide substantial assistance. With their extensive knowledge base, LLMs can evaluate software code and explain its operation. In the case of reverse engineering, LLMs can help dissect complex obfuscated code and provide insights into its functionality and purpose. By understanding how a piece of malware or exploit operates, analysts can develop effective countermeasures to protect their systems and networks, while also saving time during the investigation.

For example, VirusTotal's[8] integration of the Sec-PaLM model now generates natural language summaries of code snippets, enabling analysts to identify potential vulnerabilities more quickly. Similar features are likely to be integrated into many reverse engineering, sandboxes, and analysis tools in the future.

4.3 Automating threat intelligence

LLM transformers can greatly enhance the process of generating threat intelligence reports by automating the collection, analysis, and summarisation of relevant data. This not only saves time and effort, but it also ensures that the information presented to cybersecurity teams is accurate, up-to-date, and easily digestible. Armed with this intelligence, defenders can make more informed decisions and take proactive measures to protect their systems.

[6] Microsoft Security Copilot
[7] New wave of AI powered capabilities
[8] VirusTotal Code Insight

4.4 Coding and documentation

LLM transformers have already shown their impact on the field of software development. These AI models assist developers with code completion, bug detection, security vulnerability identification, and automatic documentation. By suggesting secure code snippets and identifying flaws early, LLMs help accelerate the development process and improve security.

Tools like GitHub Copilot X[9] or Amazon CodeWhisperer[10] are enhancing developer productivity and integrating security best practices into workflows. While not replacing human developers, these tools provide significant support, especially in preventing common vulnerabilities such as SQL injection or cross-site scripting.

Generative AI also automates routine tasks like writing documentation and generating unit tests, enabling developers to focus on more complex security issues. However, since LLM-generated code is based on public datasets, there is a risk of inheriting bugs or vulnerabilities, making human oversight essential to ensure secure, production-ready code.

For example, CERT-EU has deployed internal copilot tools that leverage the use of internal AI models to assist developers in writing secure code and identifying potential vulnerabilities. These tools have significantly improved the quality of code produced and reduced the time spent on manual code reviews.

4.5 Enhancing cybersecurity training

The sophisticated natural language processing capabilities of large language models are playing an increasingly important role in upskilling cybersecurity personnel. AI-powered systems now offer contextualised explanations of complex cybersecurity concepts, enabling junior staff to bridge knowledge gaps more quickly. For example, AI platforms can provide personalised guidance on identifying and responding to emerging threats. By offering real-time support, these models help less experienced team members contribute more effectively to threat analysis and mitigation, thereby enhancing the overall efficiency of security teams.

Moreover, many institutions are now offering more frequent internal training sessions to help staff better understand and properly utilise AI tools. These training programmes are designed to ensure that teams can leverage AI tools effectively, enhancing their ability to respond to cybersecurity challenges in a rapidly evolving landscape.

4.6 Content generation

Finally, one of the most evident applications of Generative AI is the creation of high-quality content across a range of domains, including the automatic generation of technical documentation, corporate communications, and presentations. The ability of AI-driven content generation platforms to understand context and produce human-like text can enhance organisations' approaches to content creation. The anticipated widespread adoption of Generative AI is largely due to its impressive capacity to save time, reduce costs, and increase overall efficiency in producing a variety of content types.

In the field of cybersecurity, Generative AI can be employed to draft post-incident analyses, summarise threat intelligence feeds, and produce tailored security advisories. It also supports the automation of incident reports, enabling security teams to focus on more strategic tasks. By streamlining the drafting process, organisations can ensure that critical information is com-

[9] GitHub Copilot
[10] Amazon Code Whisperer

municated clearly and promptly, facilitating quicker and more effective responses to emerging threats.

Similarly, AI-driven content generation can enhance the quality and efficiency of presentations. By leveraging data, Generative AI can dynamically generate visuals, suggest relevant talking points, and recommend persuasive storytelling techniques to engage audiences. This not only simplifies the process of creating presentations but also increases their overall impact and effectiveness.

Similarly to code generation examples, there are already several products either available or being rolled out that propose such AI-based enhancements and solutions. These include, for instance, Microsoft 365 Copilot[11] and Google AI-powered Workspace Features[12].

5 Deployment considerations of AI models

Generative AI models usually require significant computational resources to function effectively. Choosing the right infrastructure is crucial to balancing performance, scalability, cost, and security. There are several deployment options available, each with distinct advantages and trade-offs.

5.1 Public closed-source models – paid or free

The AI industry has seen a surge in public closed-source models offered by major tech companies. These models provide powerful AI capabilities but often function as black-box solutions, limiting user control over data processing and storage.

The dominant market model today revolves around "free" closed-source AI services, such as ChatGPT, DALL-E, Midjourney, and Google Gemini. These platforms make AI highly accessible but come with significant data privacy concerns. Their terms of use often indicate that input and output data may be stored outside the EU and could be used for further training and fine-tuning of the models. As a result, organisations handling sensitive information must assume that any data provided through these services could become public knowledge.

Additionally, the emergence of cost-effective AI models from competitors like China's DeepSeek has intensified market competition, prompting tech giants to reassess their offerings and pricing structures. While these services offer ease of use and accessibility, organisations must carefully evaluate whether the trade-offs in data sovereignty and compliance align with their operational and regulatory needs.

5.2 Locally-hosted open-source models

Deploying open-source AI models on local infrastructure has become a key strategy for organisations prioritising data control, security, and customisation. Models like LLaMA 3 and Mistral can be hosted on-premises or in private clouds, ensuring compliance with data sovereignty regulations while avoiding reliance on external providers.

However, maintaining these systems requires high-performance GPUs and skilled personnel, leading to the growing adoption of AI colocation services. These services provide access to cutting-edge computing infrastructure without the overhead of in-house data centre management.

[11] Microsoft 365 Copilot
[12] Google AI-powered workspace

As AI advances, organisations are increasingly balancing performance, cost, and security by adopting self-hosted open-source models and colocation solutions, ensuring greater control over AI applications while maintaining scalability.

5.3 Privacy-focused commercial closed-source models with specific conditions of use

In response to increasing concerns over data privacy and regulatory compliance, several tech companies have introduced privacy-focused AI services that adhere to strict data handling policies. These cloud providers now offer solutions that ensure user data remains within specified regions and is not used for training or fine-tuning models, striking a balance between leveraging advanced AI capabilities and maintaining control over sensitive data.

For organisations handling sensitive information, these services offer an opportunity to benefit from AI technologies while ensuring privacy compliance. Additionally, privacy-focused commercial models with negotiated terms of use are gaining traction. These models come with specific configurations and agreements that differ from public closed-source models, providing stronger data protection but imposing more stringent conditions for non-compliance. As these offerings become more prevalent, organisations must carefully assess the terms of service to ensure they align with privacy requirements and legal obligations within the evolving regulatory landscape.

When selecting a deployment option for Generative AI models, organisations must carefully evaluate their specific needs and privacy requirements. One promising option for Union entities is the adoption of privacy-focused commercial models with customised terms of use. These models offer stronger data protection than public closed-source solutions, but with stricter conditions that may carry serious consequences for non-compliance.

6 Risks

Risks associated with the use of Generative AI can broadly be divided into two main categories:

- Risks related to its use within an organisation.
- Risks resulting from its use by others, including malicious actors.

6.1 Risks of using Generative AI

Specific risks may arise from the potential use of Generative AI technology by the staff of Union entities. As with its benefits, the emphasis remains firmly on cybersecurity, with a particular focus on these organisations.

6.1.1 Indirect prompt-injection attacks

As Generative AI technology evolves, new possibilities and risks emerge. The recent development of various plugins and interfaces for external data sources that can be used in conjunction with some of the large language models – or even independent AI Agents – increase their capabilities but also introduce new risks.

One of these risks is the possibility of indirect prompt-injection attacks. When models are able to use external data – websites, documents, emails, etc., such external data may potentialy be under the control of malicious actors. This can allow an attacker to attempt to influence the model's output by carefully crafting their input or "prompt", often embedding hidden instructions or biases. The AI model then inadvertently generates output that could potentially spread misinformation, reveal sensitive information, or produce other undesirable outcomes. Despite

the input appearing harmless or neutral to a human observer, it can result in manipulated outputs, thus presenting significant security concerns in the application of AI technologies.

Indirect prompt-injection attacks are already occurring in the wild, with various tools and setups allowing LLMs to access external data sources being used as vectors for these attacks. Examples include hidden text in web pages, inside documents or emails that are then provided as input to LLMs by unsuspecting users.

A significant challenge is that the existing defences are not currently equipped to effectively counter these attacks. The subtlety of the manipulation makes detection extremely difficult, especially as the injected prompts often appear harmless or neutral to human observers or are not easily visible at all. While it is possible to configure the models to ignore certain types of these attacks or specific prompts, there is no obvious way to create a permanent fix. Users should be cautious when using AI tools on any input that may have been subject to malicious modification (e.g., web pages, external documents, incoming emails, etc.).

6.1.2 Disclosure of sensitive data

The use of freely available, closed-source AI language models, such as ChatGPT, poses potential risks to sensitive data submitted in user prompts. As users interact with the model, they may inadvertently input confidential or personally identifiable information (PII) while seeking assistance or answers. Since this information is typically stored to enable the model to process and generate responses, there is a risk that sensitive data could be exposed, either through data breaches or during the training of future versions of the AI models. Without proper data anonymisation and privacy safeguards in place, such information could be misused by unauthorised parties, leading to identity theft, financial fraud, or reputational damage for both individuals and organisations involved.

For instance, OpenAI's current terms of use[13] specify that while OpenAI does not use API content to improve their services, they may use non-API content (i.e., prompts and outputs from ChatGPT) for this purpose. Therefore, if confidential or sensitive data is entered as part of a ChatGPT prompt, it could eventually be exposed. OpenAI states that requests submitted via their API will be stored for 30 days[14] and not used for training. However, there is no guaranteed proof of compliance or transparency regarding OpenAI's future plans.

In the event of a cyberattack on the infrastructure of an AI language model, there is a significant risk of data leakage. Such a breach could expose sensitive and private user information, including personal details, confidential conversations, and intellectual property. The consequences of such exposure could be wide-ranging, including compromised privacy, loss of user trust, and potential legal ramifications.

Organisations must also be vigilant in how AI tools are deployed within their environments. Employees should be trained not to enter sensitive data into public models, and technical controls – such as data masking, secure API gateways, and audit logging – should be implemented where possible.

The European Commission's AI Act[15] seeks to enforce strict standards around AI deployment and data protection. By complying with its provisions, organisations can reduce the risk of unauthorised access, enhance system transparency, and demonstrate accountability in how sensitive data is handled. This regulatory framework fosters public trust in AI technologies and encourages investment in privacy-preserving solutions.

[13] OpenAI terms of use
[14] OpenAI retention policy
[15] Regulatory framework AI

6.1.3 Copyright violations

Generative AI technologies, such as text and image generation models, have raised concerns about potential copyright violations as they become increasingly adept at creating content that closely resembles human-authored works. In the realm of text generation, AI-powered tools can produce articles, stories, or even poetry, often blurring the lines between human creativity and synthetic, machine-generated output. This raises questions about the originality of the content and whether the AI system has unintentionally reproduced or closely mimicked copyrighted materials.

For instance, if a text-generation AI model creates a story that closely resembles a popular novel, the copyright holder of the original novel may claim infringement, arguing that the AI-generated work could be perceived as a derivative of their copyrighted material.

Similarly, image-generation models have the capability to create visually appealing artwork, designs, and even photorealistic images. These AI-generated images could infringe upon copyrighted visual content if they closely resemble existing works, such as paintings, photographs, or graphic designs. For example, if an image-generation AI model were to create an artwork strikingly similar to a famous painting, it could lead to copyright disputes between the original artist and the creator of the AI-generated piece. Moreover, these concerns extend to the potential appropriation of elements from multiple copyrighted works to create a new image, which could lead to multiple copyright violation claims.

In both cases, the increasing sophistication of Generative AI technologies complicates the legal landscape surrounding copyright protection, as it becomes more challenging to determine the true authorship and originality of content.

Additionally, in some instances, the models powering Generative AI technologies are known to be trained on copyrighted content without the explicit approval of the authors. This raises additional concerns, as the organisations behind these models could be held liable for potential copyright infringement. By using copyrighted material to train their AI systems, organisations may inadvertently propagate the unauthorised reproduction or adaptation of protected works, opening themselves up to potential litigation. As a result, there is a growing need for more robust and transparent content acquisition policies to ensure that the data used to train AI models is either appropriately licensed or falls under the scope of fair use.

6.1.4 False or inaccurate information

AI language models have become increasingly adept at generating high-quality text. However, these models have flaws, and the risk of providing false or inaccurate information remains significant[16]. As AI language models are trained on vast amounts of data from the internet, they are susceptible to absorbing and perpetuating the biases, misconceptions, and inaccuracies present in that data. It is also important not to confuse Natural Language Processing (NLP), which these models excel at, with Natural Language Understanding (NLU), a significant challenge in AI research. A system trained solely on form has a priori no way to learn meaning[17]. Consequently, users of these models must be aware of the potential pitfalls and exercise critical thinking when interpreting generated text.

One primary concern with AI language models is the possibility of bias. As these models learn from the data they are trained on, any biases present in the training data are likely to be absorbed and perpetuated by the model. This could manifest as gender, racial, or political biases, among others, and can lead to the generation of text that is offensive or perpetuates

[16] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
[17] Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

harmful stereotypes. In some cases, these biases may even cause the AI to provide misleading or outright false information, leading users astray and potentially reinforcing pre-existing biases.

Similarly, when generating computer code in various programming languages, Large Language Models (LLMs) often provide code containing errors or security vulnerabilities. This is primarily due to the training data these models are exposed to, which may include a diverse array of programming languages, styles, and quality levels. As a result, LLM-generated code may not always adhere to best practices or conform to the latest security standards. Additionally, these models lack the inherent ability to understand the context or specific requirements of a given task, which may lead to the production of code that is unsuitable, flawed, or even dangerous. Therefore, it is crucial for developers to carefully review and validate any code generated by LLMs before incorporating it into their projects to mitigate potential risks and ensure the safety and integrity of their software applications.

Another concern is the phenomenon of "hallucinations" where AI language models generate text that appears plausible but is entirely fabricated or lacks a factual basis. These hallucinations can occur for various reasons, such as the model trying to fill in gaps in its knowledge or attempting to provide a coherent response to an ambiguous or unfamiliar prompt. While these hallucinations can sometimes be relatively harmless, in other instances, they can lead to the dissemination of false information or contribute to the spread of misinformation.

6.1.5 Hype abuse

The rapid advancements in Generative AI technology and the surrounding hype have led to a surge in public interest and adoption. While these innovations undoubtedly offer numerous benefits and transformative potential, the excitement can also lead to possible pitfalls. With increased hype, bad actors may exploit the situation by creating fake applications or investment schemes, capitalising on the popularity of recognisable AI brand names to deceive users and achieve malicious objectives.

One such pitfall is the emergence of fake ChatGPT apps on Android and iOS platforms. These counterfeit apps, disguised as popular AI language models, may carry out invasive data harvesting activities. Unsuspecting users who download and interact with these malicious apps may inadvertently expose their personal information, including messages, contacts, and browsing history. The harvested data can then be used for various nefarious purposes, such as identity theft, targeted advertising, or even extortion. This underscores the importance of exercising caution when downloading mobile applications and ensuring they originate from trusted sources and developers.

Another potential pitfall linked to the hype around Generative AI is the creation of fictitious cryptocurrency tokens using recognisable AI brand names. Bad actors may design and market these tokens to lure in unsuspecting investors, who may believe they are investing in a promising AI venture. Once the scammers accumulate a substantial amount of funds, they may disappear, leaving the investors with worthless tokens and significant financial losses. This highlights the need for investors to conduct thorough research and due diligence before committing to any investment, particularly in emerging technologies like AI and cryptocurrencies.

6.1.6 Over-relying on technology

Over-relying on Generative AI technology presents several potential dangers that could significantly impact organisations. One key concern is the possible erosion of competence among staff. As AI systems become increasingly capable of handling tasks traditionally carried out by humans, employees may become more reliant on these technologies. This dependence could lead to a decline in critical thinking and problem-solving skills, making staff less versatile and

adaptive when faced with new challenges. Furthermore, as AI takes over routine tasks, workers may lose the ability to perform these manually, resulting in a loss of valuable expertise.

Another issue is the overconfidence in the quality of output generated by AI. Due to inherent limitations in AI models, such as the token limits that restrict the amount of information a language model can "remember", the generated content may not be as accurate, comprehensive, or contextually appropriate as users might expect. This could lead to situations where AI-generated content is accepted at face value, potentially causing misinformation or poorly informed decisions.

The over-reliance on AI technologies may also manifest as a failure to account for policy or political decisions that could limit their use. Governments and regulatory bodies are increasingly scrutinising the implications of AI on privacy, security, and social equality, as seen with regulations like the AI Act[18]. Consequently, new policies or regulations may be introduced, imposing restrictions on the development, deployment, or use of AI technologies. Organisations that become overly dependent on AI systems may find themselves ill-prepared to adapt to these changes, leading to potential operational disruptions.

Finally, as highlighted in the benefits section, when using LLM tools for programming, it is essential to remember that the generated code may contain bugs or be insecure or unsuitable. Extra care must be taken when allowing staff and contractors to use LLMs for developing applications. The emphasis should be on ensuring thorough validation and testing of the generated code to mitigate potential risks.

6.1.7 LLMs opinions, advice, and moral values

Large Language Models (LLMs), such as ChatGPT or DeepSeek, should not be relied upon for opinions, advice, or moral guidance due to the inherent limitations in their design and the nature of their training data. While LLMs are powerful AI tools, they are not human beings with emotions, life experiences, or ethical systems. Instead, they are complex algorithms built to generate humanlike text based on patterns and associations identified within vast amounts of data.

One of the primary reasons LLMs are unsuitable for providing opinions, advice, or moral guidance is that their responses are derived from the datasets used in their training. These datasets consist of vast amounts of text from a wide variety of sources, which may contain conflicting opinions, values, and perspectives. When an LLM encounters such contradictions, it may struggle to produce a coherent and consistent response. As a result, the output may seem random, as the model attempts to balance opposing viewpoints or may simply reproduce popular opinions without understanding the underlying reasons or nuances.

Furthermore, LLMs are incapable of forming independent opinions or moral judgments. They do not have the capacity to critically analyse complex issues or empathise with human emotions, both of which are essential when providing sound advice or ethical guidance. Relying on an LLM for such matters could lead to misguided or superficial conclusions that fail to address the unique complexities of a given situation.

It is therefore not surprising that, for example, Deepseek has imposed strict guardrails on the values that LLMs must reflect, aligning them with those of the Chinese Communist Party, as part of its broader efforts to enforce censorship and control over AI-generated content[19]. After all, an LLM will inevitably reflect the moral values embedded in its training data and shaped by human feedback (Reinforcement Learning with Human Feedback – RLHF). Consequently, for

[18] AI Act
[19] How DeepSeek Censorship Actually Works

any generated text intended for political purposes, it may be wise to verify whether it aligns with the general vision, policy, and strategy of Union entities.

6.2 Risks from adversarial use of Generative AI technology

Specific risks arise from the use of Generative AI technology by malicious actors. As previously mentioned, the focus remains on cybersecurity, particularly with regard to Union entities.

6.2.1 Privacy issues

Personally identifiable information (PII) can inadvertently be included in the training datasets of generative AI models when data is collected from a broad range of sources, such as websites, forums, social media, and other digital platforms. This data may not always be properly anonymised or sanitised before being used to train the models. As a result, PII — including names, addresses, telephone numbers, email addresses, or other sensitive details — could become embedded within the model's training data, potentially allowing it to be traced back to specific individuals.

When these models are deployed, there is a risk that PII could be unintentionally disclosed through generated outputs. However, malicious actors may also deliberately exploit generative AI models to extract or reconstruct sensitive information, using techniques such as prompt injection or model inversion. These attacks are designed to probe the model for private or confidential data that may have been memorised during training.

This presents a twofold privacy risk: firstly, the unauthorised disclosure or targeted extraction of sensitive information could have serious consequences for the individuals concerned; secondly, the generated content may be inaccurate or misleading, resulting in misinformation or the misidentification of individuals. Both scenarios pose significant threats to privacy, trust, and the safe adoption of AI technologies — particularly in sensitive sectors such as government, healthcare, and finance.

6.2.2 More advanced cyberattacks

Generative AI technologies could also give rise to new methods for conducting cyberattacks. As AI systems become more sophisticated, they can be increasingly exploited by malicious actors to facilitate attacks and exploit vulnerabilities in various ways.

One such method involves using AI to generate phishing content. By harnessing natural language processing and generation capabilities, cybercriminals can craft highly convincing emails, text messages, and social media posts that appear to come from legitimate sources. These AI-generated messages can be specifically tailored to target individuals, increasing the likelihood of them falling victim to the scam. Furthermore, AI can automate the process of sending phishing messages, allowing attackers to target a larger number of potential victims.

Social engineering attacks can also be enhanced by AI-generated voice and video deepfakes. These realistic forgeries can be used to impersonate executives, celebrities, or other influential figures, manipulating victims into providing sensitive information or taking actions that benefit the attacker. Deepfake technology can also be employed to create more convincing phone scams or video calls, further increasing the likelihood of a successful attack.

Additionally, AI technologies can be used to improve malware, making it more difficult to detect and more effective in its operations. For example, AI algorithms can analyse existing malware and identify patterns likely to be flagged by antivirus software. Based on this analysis, AI can

then generate new, stealthier malware variants that are harder to detect and better exploit system vulnerabilities.

AI can also facilitate cyberattacks through more efficient vulnerability detection and fuzzing. By using AI-powered tools, attackers can rapidly discover security weaknesses in software or network infrastructure, much faster than traditional methods. This allows them to identify and exploit vulnerabilities before they are patched, increasing the likelihood of a successful attack.

Furthermore, AI can be used to automate and optimise password cracking. By employing machine learning algorithms, attackers can recognise patterns in password creation and generate more effective password dictionaries, significantly speeding up the cracking process. This can drastically reduce the time it takes to gain unauthorised access to accounts, making it more difficult for security professionals to respond effectively.

Finally, the development of freely available Generative AI tools has inadvertently lowered the entry barriers for new malicious actors in the cybercrime ecosystem. With minimal technical expertise, individuals can exploit the capabilities of advanced AI models to conduct various illicit activities, such as generating phishing emails, creating realistic deepfakes, or producing fake news. This democratisation of access to powerful AI-driven tools amplifies the reach and impact of cybercrime, cyber espionage, and other malicious activities. It also poses significant challenges for cybersecurity professionals, law enforcement, and policymakers, as it allows a wider range of actors to participate in these activities.

6.2.3 Disinformation

The powerful capabilities of Generative AI models come with significant risks when misused for disinformation campaigns. These models can impersonate public figures and create highly convincing narratives, making them potent tools for spreading false and misleading information. For instance, deepfake technology allows bad actors to produce fake videos and audio clips of politicians and celebrities, manipulating their words and actions to deceive the public and sow confusion. An example of this includes DRAGONBRIDGE's attempt to use AI-generated images to discredit U.S. leaders[20]. However, such campaigns have so far seen limited success.

Generative AI models can also be employed to craft realistic disinformation campaigns that undermine trust in institutions, destabilise social cohesion, and disrupt democratic processes. For example, during election periods, a sophisticated AI-generated disinformation campaign could manipulate public discourse by disseminating false news stories, conspiracy theories, and divisive content. The consequences of such actions can be far-reaching, swaying public opinion based on lies, and eroding trust in the democratic process.

The fact that these disinformation campaigns can be pre-planned and automated exacerbates the problem significantly, allowing malicious actors to generate and disseminate false information at an overwhelming scale. This makes it extremely challenging for fact-checkers, journalists, and social media platforms to identify and counteract the spread of disinformation in a timely manner. Moreover, the speed and efficiency with which AI can produce content makes it harder for users to distinguish between legitimate and fake news, further facilitating the spread of misinformation.

6.2.4 Censorship and control

Large AI models can also be exploited by authoritarian governments to manipulate public opinion and suppress democratic processes. By using these advanced technologies to generate fake news, propaganda, and deepfake content, such regimes can craft an illusion of reality that

[20] Google disrupted over 10,000 instances of DRAGONBRIDGE activity

aligns with their interests. This disinformation can cause confusion and distrust among the public, undermining the credibility of democratic institutions and opposition leaders.

Additionally, authoritarian governments can utilise AI-powered surveillance systems to track and monitor the activities of opposition members and dissidents. By analysing vast amounts of data from social media, communications, and location tracking, these models can create detailed profiles of individuals considered threats to the regime. The authorities can then use this information to suppress dissenting voices through harassment, arrests, or other forms of repression.

7 Conclusions and recommendations

Generative AI technology has emerged as a transformative innovation with the potential to disrupt industries and reshape society. While predicting its future trajectory is challenging due to the rapid pace of technological evolution, past trends provide valuable insights. The launch of ChatGPT marked a pivotal moment, sparking widespread interest and changing the way businesses and individuals interact with AI. This prompted competitors like Google and Anthropic to release proprietary models, although these remain tightly controlled by their developers. At the same time, the rise of open-source models, such as those based on Meta's LLaMA or Mistral AI, has democratised access, allowing organisations and individuals to deploy, customise, and run AI tools independently at lower costs. These models now rival their closed-source counterparts in performance, providing privacy-conscious entities with a viable alternative for on-premises deployment.

However, this immense potential comes with inherent risks. The ability of Generative AI to create realistic content raises ethical concerns, including the proliferation of deepfakes, misinformation, and automated job displacement. Biases embedded in training data threaten to perpetuate discrimination, while the technology's rapid advancement poses challenges for regulatory frameworks. Yet, Generative AI is not magic – it builds on decades of research in machine learning, neural networks, and computational power. Sophisticated algorithms process vast datasets to produce complex outputs, with today's state-of-the-art tools merely serving as a precursor to even more capable systems.

As progress in AI remains inevitable, organisations must proactively integrate these tools into their strategies while establishing ethical guidelines and security protocols. The EU's €200 billion investment in AI infrastructure highlights the urgency of this task, striving to balance innovation with accountability. Failing to engage with this shift risks ceding competitive ground to those willing or maliciously inclined to exploit its benefits. Generative AI is here to stay, demanding a dual focus on harnessing its potential while mitigating its dangers.

7.1 Recommendations

We present recommendations to support Union entities in directing and coordinating their efforts concerning generative AI. Due to the rapidly evolving nature of this field, the recommendations are categorised into short-, medium-, and long-term measures.

7.1.1 Short-term

- **Stay informed on Generative AI developments**: Continuously monitor advancements in Generative AI, as these technologies will likely impact various aspects of your operations and workflows.

- **Invest in user awareness and training**: Promote responsible and informed usage of AI within your organisation by ensuring that staff understand both the benefits and risks associated with the technology.
- **Establish data handling policies**: Implement clear guidelines to ensure that only publicly available data (TLP: CLEAR) is used in prompts submitted to commercial large language models, especially those provided online by public AI services.
- **Explore services under Cloud Broker PaaS Contracts**: Investigate privacy-enhanced commercial AI models, such as Azure OpenAI or Mistral, which provide secure environments with regional hosting, ensuring user data stays within the EU and isn't used for further training.
- **Monitor open-source models**: Local, open-source models are progressing rapidly, with increasing potential for customisation and fine-tuning on sensitive data. Keep track of these developments for future deployment opportunities.
- **Engage with other institutions, such as CERT-EU, for expert guidance on securing your AI systems and mitigating risks related to cybersecurity and data privacy.** CERT-EU's support can help ensure that your organisation follows best practices in AI security.

7.1.2 Medium-term

- **Develop a responsible AI policy**: Establish clear internal policies to guide the ethical and responsible use of Generative AI technologies. Define acceptable use cases and implement validation processes for AI-generated outputs.
- **Plan for local infrastructure deployment**: Consider deploying local, open-source Generative AI models either on-premises or within private cloud environments. This would provide enhanced control over sensitive data while benefiting from the growing capabilities of self-hosted models.
- **Leverage the European Union's strategic AI initiatives**: With the EU committing to investments in AI and its efforts to create an ecosystem that prioritises data sovereignty and security, consider aligning your AI strategy with regional initiatives and future EU-backed cloud services to ensure compliance with privacy regulations.

7.1.3 Long-term

- **Invest in advanced AI infrastructure**: As Generative AI continues to evolve, it will be crucial to invest in scalable and resilient infrastructure to support complex AI workloads. Consider establishing partnerships with cloud providers or with other institutions to ensure that your organisation can handle increasing data and processing demands in a cost-effective and efficient manner.
- **Foster AI-driven innovation**: Encourage research and innovation within your organisation by establishing AI-driven labs or collaborating with other institutions such as CERT-EU. This can help create specialised models tailored to your specific needs, advancing your competitiveness and strategic positioning in the AI space.
- **Prepare for next-gen AI applications**: Look ahead to the next-generation AI technologies, such as AI systems with advanced reasoning, multi-modal capabilities, and autonomous decision-making. Prepare your organisation to leverage these innovations through long-term AI strategy planning, including upskilling your workforce and investing in future technologies.

8 Credits

We would like to warmly thank our colleagues from European Commission for the useful input, feedback and suggestions they have provided to improve this guidance.

9 Contact

If you have suggestions that could help improve this document, please contact us at services@cert.europa.eu. We always appreciate constructive feedback.

TLP Definition

TLP	Disclosure	Message
RED	Not for disclosure, restricted to participants only.	Recipients may not share TLP:RED information with any parties outside of the specific exchange, meeting, or conversation in which it was originally disclosed.
AMBER	Limited disclosure, recipients can only spread this on a need-to-know basis within their organisation and its clients.	Recipients may share TLP:AMBER information only with members of their own organisation.
AMBER+STRICT	Limited disclosure, recipients can only spread this on a need-to-know basis within their organisation only.	Recipients may share TLP:AMBER+STRICT information only with members of their own organisation.
GREEN	Limited disclosure, restricted to the community.	Subject to standard copyright rules, TLP:GREEN information may be distributed with peers and partner organisations within their sector or community, but not via publicly accessible channels.
CLEAR	Disclosure is not limited.	TLP:CLEAR Recipients can spread this to the world, there is no limit on disclosure.