CS 153: GENERATIVE AI AND HACKING

Donovan Jasper
Department of Computer Science, Department of Music
Stanford University
djasper@stanford.edu

Eliot Jones
Department of Computer Science
Stanford University
ekj@cs.stanford.edu

Edward Adams
Department of Computer Science
Stanford University
edward27@stanford.edu

ABSTRACT

GenAI offers powerful tools for both enhancing cybersecurity defenses and enabling sophisticated cyberattacks. Its capabilities in threat detection, behavior analysis, and incident response can significantly strengthen an organization's security posture. However, in addition to introducing its own class of vulnerability, GenAI also facilitates automated vulnerability exploitation, polymorphic malware creation, and advanced phishing attacks. To manage these risks, organizations must adopt comprehensive strategies, including meticulous data curation, robust reinforcement learning, rigorous testing, and prompt filtering mechanisms. Regulatory frameworks like the EU's AI Act and NIST guidelines are good steps, but the primary responsibility lies with the developers of these technologies. By implementing best practices and adhering to regulations, organizations can leverage the benefits of GenAI while minimizing its risks, improving the navigation of the dual-edged nature of this transformative technology.

1 INTRODUCTION

1.1 SECURITY TOPIC

The field of cybersecurity is continuously evolving with new threats and vulnerabilities emerging as technology advances, especially in fast-paced, profit-driven entrepreneurial environments. Hacking, within the cybersecurity domain, refers to the practice of exploiting vulnerabilities in computer systems and networks to gain unauthorized access to data, disrupt operations, or achieve other malicious objectives. This activity encompasses a wide range of techniques, including but not limited to phishing, malware deployment, SQL injection, and brute force attacks (1).

Generative AI (GenAI) has the potential to revolutionize cybersecurity, offering new tools for both defense and attack. This paper explores the role of GenAI in penetration testing and vulnerability assessment. Our focus will be on its application in red team-based testing. By examining current research with practical examples, we aim to highlight the potential and limitations of AI-driven approaches in this domain.

1.2 RELEVANCE

We selected this topic because recent studies indicate that AI-driven cyberattacks are increasing in frequency and sophistication. There are several instances where AI-assisted attacks have significantly impacted organizations. One notable instance is the massive AI-controlled botnet assault on

WordPress sites (2). Additionally, a 2023 report by CFO found that 75% of cybersecurity experts saw an uptick cyberattacks that year and 85% of those professionals believe that hackers' use of generative AI was resposible for the increase in attacks (3). According to Cybersecurity Ventures, cybercrime is predicted to cost the world $10.5 trillion annually by 2025, up from $3 trillion in 2015, reflecting the growing scale and impact of these threats (4). Given this alarming trend, we believe it is valuable to understand the current state of GenAI technology and its implications for cybersecurity. The rest of this background section will focused on the nuanced and evolving capabilities of GenAI due to the clear and present nature of the threats and consequences evidenced by these findings.

1.3 Emerging Technology and Threat Models

This portion of the background section will rely primarily on research examples because they tend to offer greater transparency and broader perspective compared to industry reports. Although research examples can sometimes present results in a more favorable light, we believe we have addressed this potential bias effectively. Industry capability reports often use vague and buzzword-laden language. For instance, claims such as "By using AI and machine learning, we can improve the efficiency and accuracy of our penetration tests, prevent malicious hackers from gaining access to computer systems, and help our clients to stay ahead of the latest threats and vulnerabilities" are common. While this claim later includes a singular specific example, a trivial TryHackMe box, we believe that industry reports tend to lack detailed evidence and focus more on marketing. Therefore, our background aims to provide a clearer, evidence-based perspective on the capabilities and limitations of GenAI in cybersecurity before diving into more industry-driven applications (5).

1.3.1 Red Team-Based Research Background

Red team-based testing involves simulating real-world attacks to evaluate the security posture of an organization. This type of testing goes beyond traditional vulnerability assessments by actively exploiting vulnerabilities to understand the impact of potential attacks. According to the paper 'A Systematic Literature Review and Meta-Analysis on Artificial Intelligence in Penetration Testing and Vulnerability Assessment', a significant portion of AI research leverages red team-based approaches to assess AI's effectiveness. Specifically, 31 papers were reviewed, with 67.74% (21 papers) being conference papers and 32.26% (10 papers) being peer-reviewed journal papers (6). Some of these papers also evaluate red-teaming the models themselves, which represents a distinct class of vulnerability and will be primarily covered in the 'Mitigations' section.

In 'Getting Pwn'd by AI: Penetration Testing with Large Language Models', the use of large language models (LLMs) in assisting penetration testers was explored. An interesting case involved the AI suggesting the use of:

```
sudo /usr/bin/perl -e 'exec "/bin/sh";'
```

to gain root access. This command opened a root shell because 'perl' was allowed to run as root without a password. While this method is straightforward and something an automated tool or an experienced hacker could achieve relatively easily, it showcases the potential of AI in identifying and exploiting such vulnerabilities. However, it also highlights the current limitations, as these techniques are well-known and relatively simple (7).

1.3.2 CTF-Based Research Background

Capture The Flag (CTF) competitions are a crucial part of cybersecurity education and skill development. They simulate real-world hacking scenarios, enabling participants to practice and improve their skills in a controlled and legal environment. CTF competitions can be categorized into two main types: Jeopardy-style and Attack-Defense style. In Jeopardy-style CTFs, participants solve a variety of tasks from different categories to earn points. These tasks can include cryptography, steganography, binary exploitation, reverse engineering, and web security. In Attack-Defense CTFs, teams must protect their own systems while simultaneously attacking the systems of their opponents. This type of CTF requires participants to apply a wide range of skills in real-time, including system hardening, network security, and incident response.

Recent "Hacking with AI" papers often use Jeopardy-style CTF competitions as benchmarks for evaluating AI capabilities. However, many of these benchmarks use relatively simple or outdated CTFs, which can contaminate the tests as the AI might have been trained on that data. For example:

- The 'Language Agents as Hackers' paper used 100 PicoCTF problems, which were completed by two authors with "rudimentary computer security" skills in six hours. This showcases the current capabilities of GenAI in handling relatively simple and educational CTF problems (8).
- The 'Advanced AI Evaluations at AISI' paper frequently uses PicoCTF and CSAW, indicating a reliance on these well-known and simpler CTF challenges. (9).
- The 'CyberSecEval 2' results of a generator script were also included. Their buffer overflow test suite repeatedly uses the same code with different buffer sizes, failing to capture the intricacies and variability needed for a thorough AI evaluation. (10).
- The 'How Far Have We Gone In Vulnerability Detection Using Large Languague Models' paper uses the BUUOJ CTF, which is five years old, running the risk of contamination. (11) (12).

There appear to be more benchmarks using Jeopardy-style CTFs likely because these environments are easier to control and build. Jeopardy-style CTFs, often include categories of particular note for hacking such as web, crypto, pwn, and reverse engineering (rev), but these may not always capture the full complexity of real-world scenarios. It would be beneficial to see more research conducted in Attack-Defense CTFs or lab environments to better evaluate AI capabilities in different realistic and varied scenarios.

1.4 Background Summary

Recent research has shown that GenAI is at a stage where it can assist in solving CTF problems and even hack simple machines, showcasing its potential in educational, competitive, professional, and adversarial settings. However, the use of outdated or trivial benchmarks limits the assessment of GenAI's true capabilities in cybersecurity. It would be good to implement more challenging and up-to-date test environments to provide a clearer picture of what GenAI can achieve in real-world scenarios.

2 Interaction Risk Analysis

The interaction between GenAI and hacking presents both opportunities and risks. This section explores offensive and defensive interactions, and the resultant risks created or mitigated by these interactions.

2.1 Offensive Interactions

GenAI can significantly enhance the offensive capabilities of hackers by automating various aspects of cyberattacks. The most prominent new capabilities include:

- **Vulnerability Identification and Exploit Generation:** GenAI can rapidly analyze code and identify vulnerabilities, generating exploits that can be used to attack systems. This capability reduces the effort required to find and exploit weaknesses, making it easier for attackers to launch sophisticated attacks (13). For instance, tools like ZeroFox's FoxGPT and Tenable's ExposureAI leverage GenAI to analyze large datasets and generate insights about vulnerabilities and attack paths (14)(15).
- **Polymorphic Malware Creation:** GenAI can create polymorphic malware that change its code with each new infection. This type of malware can evade traditional security heuristics that rely on the assumption that malware code does not change across infected devices (13; 16).
- **Mass Sophisticated Phishing Attacks:** GenAI can create convincing spear phishing emails en masse. Such emails lack the spelling and grammatical mistakes of traditional

mass-phishing attacks and recent research suggests GenAI outperforms human attackers by as much as 60% in spear-fishing campaigns (17; 18). Crucially, it allows attackers to automate what is normally a human process, dramatically increasing the potential volume of high-quality phishing attacks.

- **Deepfakes and Social Engineering:** GenAI can create convincing audio and video deepfakes that can be used for social engineering attacks (19). These deepfakes can impersonate individuals, manipulate public opinion, or deceive targets into divulging sensitive information. An estimated 90% of all attacks include a social engineering component, making this a worrying threat vector (20).

These offensive applications of GenAI significantly enhance the capabilities of cybercriminals and make it easier to conduct large-scale, complex attacks with minimal human intervention.

2.2 Defensive Interactions

On the defensive side, GenAI can be a useful asset for cybersecurity teams. The most promising applications of GenAI for cybersecurity are:

- **Threat Detection:** The algorithms used to train potentially malicious GenAI rely on large datasets which can be retailored for threat classification (21). Security teams can identify threats with greater accuracy with the aid of GenAI (22).
- **Behavior Analysis:** GenAI can generate models of normal user or network behavior and identify deviations that may indicate a security breach. Security professionals can detect potential threats and take appropriate measures to prevent security incidents by analyzing such models (13).
- **Malware Identification and Analysis:** The same GenAI used by cybercriminals to create malware can be used by cybersecurity experts to create malware to train against. Such use can help security teams discover new attack vectors and better identify and defend against AI-generated malware (23).
- **Incident Response:** Generative AI can automate the initial steps of incident response, such as categorizing incidents based on severity and recommending mitigation strategies. This can reduce the time it takes to respond to security incidents and is especially helpful to organizations that have not invested in a robust, dedicated security team (13).

2.3 Risks Created by GenAI Interactions

GenAI offers significant benefits but its use also introduces new concerns. Among the avenues of interaction identified in the first section, the most relevant new risks are:

- **More Attacks:** Automation of what can otherwise be a labor-intensive endeavor means the volume of cyberattacks is likely to increase. (15). This is true along both software and social vectors and is likely to accelerate as the barrier to entry continues to lower.
- **New Attack Vectors:** The analytical ability of GenAI and the threat of polymorphic malware provide new risks for previously-unseen threat vectors. This is also true of deepfakes, where considerably less work has been done to protect against than traditional forms of social engineering. The number of attack vectors has increased and is projected to continue to rise as attackers find new ways to use GenAI (24).
- **Adversarial Attacks on AI Systems:** Many enterprises seek to employ their own GenAI for business purposes. This introduces new problems not found in typical security settings. Cybercriminals can launch adversarial attacks on AI systems by employing tactics such as data poisoning or model inversion attacks, which can disrupt or manipulate GenAI models (25).
- **Dependency on AI:** Over-reliance on AI-driven tools can lead to complacency among security teams. If AI systems fail or are compromised, organizations may find themselves unprepared to deal with cyber threats effectively (21).

2.4 Risks Mitigated by GenAI Interactions

Despite these challenges, GenAI also mitigates certain risks:

- **Enhanced Threat Detection:** Automated usage of GenAI can help security teams identify and respond to threats more quickly, reducing the potential damage caused by cyberattacks. This is especially true of conventional cyberattacks, which the model is more likely to be trained on (22).
- **Proactive Security Measures:** The most effective way to mitigate a breach is to prevent it from happening. GenAI provides a way to automate creative cybersecurity training through malware generation and penetration testing (21).
- **Improved Incident Response:** GenAI can streamline the incident response process, enabling faster and more effective mitigation of security incidents (13). While it is not a substitute for a robust security team, smaller organizations might not employ such professionals and GenAI has utility in assisting a team that might otherwise lack expertise with cybersecurity.

The interaction between GenAI and hacking presents a complex landscape of opportunities and risks. By understanding these dynamics and implementing appropriate mitigation strategies, organizations can harness the benefits of GenAI while minimizing its potential threats.

3 Mitigations

As discussed in the previous section, GenAI is potentially a very powerful tool for democratizing access to hacking. Its generative capabilities allow it to provide technical support to bad actors with malicious intents who might not have the requisite hacking capabilities. Further, while LLMs are proven to be excellent when it comes to analyzing and writing code, other models that are trained to produce audio and visual content are already proving to be the next generation of offensive social engineering and phishing tools. Deepfaked images, video, and spoofed audio are becoming increasingly convincing, even to the well-trained eye. As with any toolset, the creators must be conscious of all use cases – for good and bad – when choosing to publish said tool to the public. As a result, the majority of the responsibility for mitigating these risks falls on the shoulders of the organizations that create these tools, as well as national or international governing bodies for regulating the GenAI industry.

3.1 Organizational Level

At an organizational level, approaches to securing LLMs or other GenAI models come in three phases: the pre-model phase, the model phase, and the post-model phase. The pre-model phase denotes the period before the training of the model, when datasets are being curated, created, and cleaned prior to being used for training. The model phase represents the combination of the pretraining and fine tuning phases, and the post-model phase considers safeguards on how the outside world interacts with the model, such as through prompt filtering. There is also a fourth category which has made its way into the conversation recently, after Anthropic's findings on the interpretability of some of their model's features – this will be called the "model understanding" phase.

3.1.1 The Pre-Model Phase

Different organizations take different approaches to how they mitigate the potential harmful impacts of their GenAI models, and these differences are perhaps the most clear during this phase. For example, prior to training their DALL-E model, OpenAI removed "the most explicit content from the training data,"(26) thereby, according to them, minimizing the model's exposure to, and knowledge about, those topics. However, unlike OpenAI, it is unclear if many of the other most popular text-to-image generation models utilize the same sort of filtering. For example, Stable Diffusion utilizes (27) subsets of the LAION-5B dataset, a dataset of images scraped from the web. However, as the curators of LAION stated in their release (28), "Be aware that this large-scale dataset is uncurated. Keep in mind that the uncurated nature of the dataset means that collected links may lead to strongly discomforting and disturbing content for a human viewer." Even worse, a study (29)

by Stanford's Internet Observatory found that the dataset contained samples of child sexual abuse material, prompting LAION to take down the datasets.

Compared to the current state of the art LLMs, these text-to-image generation models are relatively small. Llama-3, for example (30), was pretrained on over 15 trillion tokens, compared to the around 2.3 billion that Stable Diffusion was initially trained on. This means that mitigations in this phase are much more feasible for smaller models that are used to generate images and video than they are for large language models. In this phase, OpenAI's method of rigorously filtering the dataset before training represents the industry best practice for smaller models.

3.1.2 THE MODEL PHASE

LLMs and other GenAI models often go through an initial stage of pretraining, followed by a fine tuning round, before being put into production. It is during this phase, especially in the fine tuning mode, where the most significant safety strides are made by most organizations. For example, in their technical paper (31) on the release of GPT-4, the authors describe their RLHF (reinforcement learning with human feedback) pipeline, in which humans are inserted into the training pipeline in order to guide the model towards more desirable outputs. They also utilize a system of GPT-4 classifiers which are used to reward the model for refusing to generate unsafe outputs. More recently, in their release (30) of Llama-3, Meta describes a similar process, whereby they utilize a "red teaming approach" consisting of both human experts, as well as some automated methods, to try to generate prompts that will elicit an undesirable or unsafe response from the model.

However, finetuning, when left to the hands of the users, can prove to circumvent the safeguards put in place by these organizations. This (32) paper by a group of researchers from Stanford, Princeton, Virginia Tech, and IBM, proved that models' built-in safeguards can be circumnavigated by as little as 10 adversarial training examples. Yet another conclusion made in the same paper was that even fine-tuning with alternative, commonly-used datasets, can reduce the efficacy of these safeguards.

3.1.3 THE POST-MODEL PHASE

The post-model phase can be further subdivided into two parts. Firstly, once a model has been trained and fine tuned, organizations typically run a suite of tests to determine the risks present. The best such example currently is Meta's CyberSecEval2 (33), which is their own benchmark used to determine the risks posed by their in-house models. This benchmark covers five categories: "insecure coding," "cyberattack helpfulness," "prompt injection," "vulnerability identification and exploitation," and "code interpreter abuse." These categories each include a series of prompts meant to test the model's resilience to prompts posed by actors with malicious intents. As mentioned in the introduction, another common form of benchmarking LLMs specifically is by using CTFs to evaluate the model's ability to solve hacking challenges. However, at present the current state-of-the-art in this category is a collection of PicoCTF problems, which have been mentioned previously as purely educational in nature, aimed at middle and high school students, thus highlighting the lack of a true benchmark in this case.

Further, any GenAI platform will come with built-in safeguards on the prompting side. DALL-E, for example, will refuse to create images of public figures in a negative light, just like GPT-4, Llama-3, and all of the other premier LLMs have a prompt filter that prevents the models from even getting access to prompts deemed to be malicious in their intent. These safeguards are incredibly important when it comes to preventing users from gaining hacking knowledge, but malicious knowledge in general, as they are the first line of defense for these models.

3.1.4 THE MODEL UNDERSTANDING PHASE

In light of Anthropic's breakthrough (34) research on LLM security and bias, this fourth category also exists as an important mitigation strategy. The researchers found a way to see which features were present as a model was generating responses, which has massive impacts on LLM safety and security. One such example they provided was the topical feature "backdoors," in reference to backdoors in systems, which can be found in responses to questions or prompts about hidden cameras. However, these features are largely controllable with what they called "clamping," or a method of retroactively manually tuning the weights of these features after training. Tuning down some of these threatening parameters resulted in less biased or hateful speech.

3.2 Governmental Regulation

At the regulatory level, there have been trends towards stricter guidelines for the ethical use of GenAI within the cybersecurity field. The European Union, for example, passed the AI Act (35), which seeks to specifically regulate systems which fall into two of their predefined categories: systems that create an "unacceptable risk," and systems which have "high-risk applications." Systems that create "unacceptable risk," such as social-scoring systems used in China, are banned, while systems with "high-risk applications," like ones that scan C.V.s during a hiring process, are highly regulated. Further, the National Institute of Standards and Technology (NIST) in the United States is currently working on developing a framework to ensure the safety and reliability of AI systems (36). However, it is unclear whether or not these types of regulations will prove to be strict enough, due to their lack of specificity and non-technical language.

3.3 Best Practices

While the importance of governmental regulation cannot be understated, the responsibility for mitigating these risks ultimately lies with the organizations who create GenAI models. The best methods for mitigation at the organizational level are:

- Removing harmful content from the initial datasets, in order to limit model exposure from the beginning.
- Utilizing RLHF during the fine-tuning stage to help guide models towards more user-friendly outputs.
- Intensive benchmarking and red-teaming after the first two phases, in order to ensure the safety and security of the model in the real world.
- Implementing safeguards into the user interaction component, in order to prevent malicious actors from prompting the model into outputting something against company or governmental policy.
- Understanding the specifics of the model's architecture and response patterns, in order to potentially catch any biases which might cause dangerous responses.

4 Citations

References

[1] Fortinet. What is hacking?, 2024. Accessed: 2024-06-09.

[2] Sam Bocetta. Has an ai cyber attack happened yet?, 2022. Accessed: 2024-06-09.

[3] Matthew Heller. Cybersecurity attacks to increase due to generative ai, say 85Accessed: 2024-06-09.

[4] Steve Morgan. Cybercrime to cost the world $10.5 trillion annually by 2025, 2021. Accessed: 2024-06-09.

[5] Red Sentry. ChatGPT, AI, and Penetration Testing. `https://www.redsentry.com/blog/chatgpt-ai-and-penetration-testing`. [Accessed 09-06-2024].

[6] D.R. McKinnel, T. Dargahi, and A. Dehghantanha. A Systematic Literature Review and Meta-Analysis on Artificial Intelligence in Penetration Testing and Vulnerability Assessment. *Computers and Electrical Engineering*, 2019.

[7] Andreas Happe and Jürgen Cito. Getting Pwn'd by AI: Penetration Testing with Large Language Models. *ESEC/FSE 2023*, 2023.

[8] John Yang, Akshara Prabhakar, Shunyu Yao, Kexin Pei, and Karthik R Narasimhan. Language agents as hackers: Evaluating cybersecurity skills with capture the flag, 2023. Accessed: 2024-06-09.

[9] Technical staff at AISI. Advanced ai evaluations at aisi: May update, 2024. Accessed: 2024-06-09.

[10] Sahana Chennabasappa Yue Li Cyrus Nikolaidis Daniel Song Shengye Wan Faizan Ahmad Cornelius Aschermann Yaohui Chen Dhaval Kapil David Molnar Spencer Whitman Joshua Saxe GenAI Cybersec Team, Manish Bhatt. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models, 2024. Accessed: 2024-06-09.

[11] Zeyu Gao, Hao Wang, Yuchen Zhou, Wenyu Zhu, and Chao Zhang. How far have we gone in vulnerability detection using large language models, 2023. Accessed: 2024-06-09.

[12] Beijing University of Posts and Telecommunications. Buuctf online judge, 2024. Accessed: 2024-06-09.

[13] Palo Alto Networks. What is the role of ai in threat detection?, 2024. Accessed: 2024-06-09.

[14] Tenable. Generative artificial intelligence for exposure management, 2024. Accessed: 2024-06-09.

[15] ZeroFox. How generative ai is changing the cyber threat landscape, 2024. Accessed: 2024-06-09.

[16] MicroAge. Confronting the next wave of cyber threats: The rise of ai-generated polymorphic malware, 2024. Accessed: 2024-06-09.

[17] IRONSCALES. Phishing simulation testing & training, 2024. Accessed: 2024-06-09.

[18] Bozkaya A. Lim, J. Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing and Defending against AI-Generated Attacks. https://i.blackhat.com/USA21/Wednesday-Handouts/US-21-Lim-Turing-in-a-Box-wp.pdf. [Accessed 09-06-2024].

[19] UST. Embracing generative ai in cybersecurity: A guide for professionals, decision-makers, and developers, 2024. Accessed: 2024-06-09.

[20] PurpleSec. Why social engineering works, 2024. Accessed: 2024-06-09.

[21] McKinsey & Company. A gen ai risk assessment, 2024. Accessed: 2024-06-09.

[22] Bain & Company. Artificial intelligence insights, 2024. Accessed: 2024-06-09.

[23] Ambika Choudhury. 8 ways generative ai can enhance cybersecurity, 2024. Accessed: 2024-06-09.

[24] Alexander Puutio. What ceos need to know about ai and cybersecurity in 2024, 2024. Accessed: 2024-06-09.

[25] Marcus Comiter. Attacking artificial intelligence: Ai's security vulnerability and what policymakers can do about it, 2019. Accessed: 2024-06-09.

[26] OpenAI et al. DALL-E 2. https://openai.com/index/dall-e-2/. [Accessed 09-06-2024].

[27] Andy Baio. Exploring 12 million of the images used to train stable diffusion's image generator, 2022. Accessed: 2024-06-09.

[28] LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS — LAION — laion.ai. https://laion.ai/blog/laion-5b/. [Accessed 09-06-2024].

[29] Largest Dataset Powering AI Images Removed After Discovery of Child Sexual Abuse Material — 404media.co. https://www.404media.co/laion-datasets-removed-stanford-csam-child-abuse/. [Accessed 09-06-2024].

[30] Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. https://ai.meta.com/blog/meta-llama-3/. [Accessed 09-06-2024].

[31] OpenAI et al. Gpt-4 technical report, 2024.

[32] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

[33] Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models, 2024.

[34] Megan Crouse. Anthropic's Generative AI Research Reveals More About How LLMs Affect Security and Bias — techrepublic.com. https://www.techrepublic.com/article/anthropic-claude-large-language-model-research/. [Accessed 09-06-2024].

[35] EU Artificial Intelligence Act — Up-to-date developments and analyses of the EU AI Act — artificialintelligenceact.eu. https://artificialintelligenceact.eu/. [Accessed 09-06-2024].

[36] National Institute of Standards and Technology — nist.gov. https://www.nist.gov/. [Accessed 09-06-2024].