Genesis Grant

CTEC 402

In Class Assignment 6

**Part 1 (Research Paper from Outline) :**

Artificial Intelligence has become an integral part of our daily lives, from virtual assistants to decision-making systems in healthcare, education, and beyond. As AI becomes more advanced, its potential for both good and harm grows. One pressing concern is the concept of "jailbreaking" large language AI models. Jailbreaking refers to bypassing the built-in safeguards of these models to unlock responses they are designed to block, such as harmful or unethical content. In this essay, I’ll explore the dangers of jailbroken AI systems, including their misuse for spreading misinformation, breaching user privacy, and supporting malicious actions. Understanding these risks is vital for ensuring AI remains a safe and trustworthy tool.

Jailbreaking large language models exposes significant vulnerabilities in their design, as seen in platforms like ChatGPT, Google Bard, and Claude by Anthropic. Research, including studies like the MASTERKEY framework, has revealed that even with strict defenses, AI models can be manipulated into generating harmful outputs through carefully crafted prompts. My study found that ChatGPT is particularly susceptible to such manipulation, providing increasingly detailed responses to harmful queries after repeated probing. Google Bard showed moderate resistance, while Claude stood out for its robust ethical framework, effectively resisting jailbreak attempts. These findings highlight the varied approaches companies take to safeguard their systems and the challenges they face in preventing malicious exploitation. Despite some models demonstrating resilience, the overall vulnerabilities call for more comprehensive mitigation strategies to address this growing issue.

Addressing the risks of jailbroken AI models requires a multi-faceted approach, including better input sanitization, ethical framework development, and cross-model collaboration to strengthen defenses. While Claude by Anthropic shows promise with its ethical protections, other models like ChatGPT still need stronger security measures. As AI technology continues to evolve, ongoing research and innovation are critical to closing these gaps. By focusing on robust safeguards and ethical standards, we can ensure AI systems remain reliable and are used responsibly in society. Protecting against jailbreaking is not just about enhancing technology—it’s about securing a future where AI serves humanity safely and effectively.

**REFERENCES:**Deng, G., Liu, Y., Li, Y., Wong, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., & Liu, Y. (n.d.). Automated Jailbreaking of Large Language Model Chatbots. MASTERKEY.https://arxiv.org/pdf/2307.08715

Part 2.1: How to restrict user access:

Go to a folder and navigate to properties tab

A screenshot of a computer

Description automatically generated

Select Security and select a group/user name

A screenshot of a computer

Description automatically generated

Use ‘PERMISSIONS FOR ’ tab to customize permissions control for specific users

A screenshot of a computer

Description automatically generated

Part 2.2: Show how to Microsoft Defender Application Control works:

To download Microsoft Defender, go to the Windows Settings, click on "Update & Security," and then select "Windows Security." Open the "Virus & Threat Protection" section and it will automatically install if it's not already on your system. Microsoft Defender should run in the background by default, protecting your device from threats. To use it, simply open Windows Security, go to "Virus & Threat Protection," and click "Quick Scan" to check for issues. You can also customize settings by clicking on "Manage Settings" under Virus & Threat Protection Settings.

Microsoft Defender helps manage internet access by monitoring network activity and blocking harmful applications. It uses rules to allow trusted apps and prevent unknown or dangerous ones from connecting to the internet. You can manage these settings by adjusting the firewall and network protection options within the app.

Part 2.3: Show how to block/allow applications from accessing Internet in Windows 11 Firewall:

To block or allow applications from accessing the internet in Windows 11 Firewall, go to the Start menu and search for "Windows Defender Firewall." Click on "Advanced Settings" on the left side of the window. In the "Windows Firewall with Advanced Security" window, click on "Outbound Rules" to block or allow outgoing traffic, or "Inbound Rules" for incoming traffic. Select "New Rule" on the right, choose "Program," and browse to the app you want to control. Choose whether to allow or block the connection for that app and finish the setup.

This method helps manage which apps can connect to the internet by creating rules in the firewall. You can customize the rules to either permit or block specific applications based on security needs.

Part 2.4: Show how to Whitelist Apps with how Applocker works:

To whitelist apps using AppLocker in Windows 11, first, open the Start menu and search for "Local Security Policy." In the "Local Security Policy" window, navigate to "Application Control Policies" and click on "AppLocker." Right-click on "Executable Rules" and select "Create New Rule." Choose "Allow" and select the application you want to whitelist by browsing to its location or selecting a publisher. Finish the setup to allow the app to run without restriction.

AppLocker works by creating rules that define which apps are allowed or denied based on file location, publisher, or hash. By setting up these rules, you can control which apps are trusted to run on your system, effectively preventing unauthorized or potentially harmful apps from executing.

Part 2.5: Show to use Macrium reflect to image disk:

To create a disk image with Macrium Reflect, first, download and install the program from the official website. After installation, open Macrium Reflect and select the disk you want to image from the main window. Click on "Create an image of the partition(s) required to backup and restore Windows," then select a destination folder where the image will be saved. Click "Next" and configure any additional options if needed, then click "Finish" to start the imaging process.

Macrium Reflect creates a snapshot of the selected disk, saving it as an image file for backup or restoration purposes. This image can be used to restore your system to its exact state in case of failure or data loss.