Sample Generative AI Security Policy

Generative AI and Large Language Models Protection and Usage Policy

Purpose

The purpose of this policy is to define controls and best practices for protecting data and privacy, managing change, and developing software with a focus on generative AI and Large Language Models (LLMs). This is critical in mitigating potential risks and ensuring alignment with relevant laws, regulations, and industry standards.

Scope

This policy applies to all individuals and entities involved in the usage, development, management and oversight of generative AI and LLMs within the organization.

Definitions

Large Language Model (LLM) - type of deep learning model designed to understand and generate human-like text.

Hallucinations - refers to instances where the model produces outputs that are not grounded in its training data, meaning it generates information that might be entirely fabricated or not based on factual knowledge.

Biases - Bias in AI refers to systematic and unfair discrimination displayed by a model due to the data it was trained on or the way it was designed. Biases can manifest in various forms, like racial, gender, cultural, or other societal biases, leading the AI to make decisions or produce outputs that are skewed, discriminatory, or unfair.

Traceability - refers to the ability to track back how the AI model was developed, trained, and deployed, providing transparency and accountability in the AI lifecycle.

Foundation Models - Foundation models are large-scale models that serve as a base on which further models or applications can be built. These models are trained on broad data and then fine- tuned or adapted for specific tasks or applications.

Prompt - a prompt is the input text or message that is given to the AI model to initiate a response. It is the starting point of the conversation or the question that you ask the model. The AI model processes this prompt and generates a response based on its training and the patterns it has learned from the data it was trained on. The quality and relevance of the response often depend on the clarity and specificity of the prompt.

Requirements

1. **Approval of AI Tools**

To ensure that AI tools align with organizational standards and to safeguard against potential risks to security, AI tools must be approved, inventoried, and used in a secure manner.

a) Employees must submit any generative AI tool or platform intended for use for approval and review before use.
b) The organization will conduct periodic inventories to detect and map all generative AI tools currently in use, ensuring alignment with approved application lists.
c) Employees are prohibited from using unapproved AI tools or platforms for any company-related activities.
d) Create an awareness program to educate employees about approved AI tools, the dangers of unapproved or malicious AI tools, and the necessity of adhering to approved tools only.

2. Safe usage of Consumer AI products

Establish a framework that ensures the secure, ethical, and responsible use of AI within the organization. All usage of Consumer AI generators such as ChatGPT, Bard, and Bing Chat within the organization must be handled securely.

a) Educate employees on the standards of prompt creation and safe use of generative AI services.
b) Prohibit the use of sensitive or private data in input prompts.
c) Employees who generate content with generative AI should follow ethical guidelines.
d) Employees are prohibited from generating content that could be used to commit fraud, crime, impersonation, or harm to a person or the company.

3. Generated Content Usage

To ensure the reliability and accuracy of content produced by Consumer AI products within the organization, mitigating potential misinformation, and ensuring alignment with the organization's standards of quality, all content generated undergo a review and validation process to confirm accuracy and appropriateness before distribution or use. This includes text, code, video, or audio.

a) AI-generated content is proofed before use.
b) AI-generated Content is verified for factual accuracy.
c) AI-generated Content is checked for biases.
d) AI-generated Content is labeled as to its origin.

4. Integrating AI into a company product or process

Integrating AI into the company's products or processes should be done in a manner that ensures that the AI aligns with organizational objectives, respects ethical standards, and adheres to industry best practices.

a) All models utilized to store or transfer data are classified based on the data processed.
b) Access to data models is strictly controlled.
c) The usage of private AI models is monitored for both input and output.
d) Data is anonymized before being input into the data model.
e) Data output is monitored for hallucinations.
f) Use only certified and non-vulnerable open-source models or secured foundation models.
g) Model training data is vetted or fine-tuned.