OWASP Top 10 for LLMs: Protecting Large Language Models with CalypsoAI

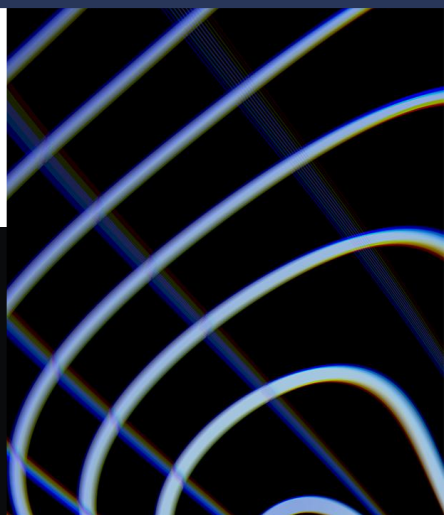
CALYPSOAI

Introduction

As cyber threats evolve, the emergence of Large Language Models (LLMs) has opened up a new attack surface that organizations must urgently address. The consequences of a security breach in an LLM are substantial, ranging from loss of sensitive data and eroded stakeholder trust to significant financial and legal repercussions. Such incidents can also damage an organization's reputation, affecting customer loyalty and future business. Therefore, addressing LLM vulnerabilities is essential not only for technical security but also for maintaining overall operational integrity and financial health in a digitally evolving world. Ultimately, it is by effectively addressing these risks that organizations can pave the way for the successful deployment and utilization of such models.

The Open Web Application Security Project (OWASP) is a vital resource for organizations facing these vulnerabilities.

Version 1.1 of the OWASP Top 10 for Large Language Model Applications highlights potential security risks encountered when deploying and managing LLMs.



In this white paper, we explore how CalypsoAI's wraparound LLM security and enablement solution, addresses each of the 10 LLM application security concerns raised by the OWASP standards.

LLM01 - Prompt Injections

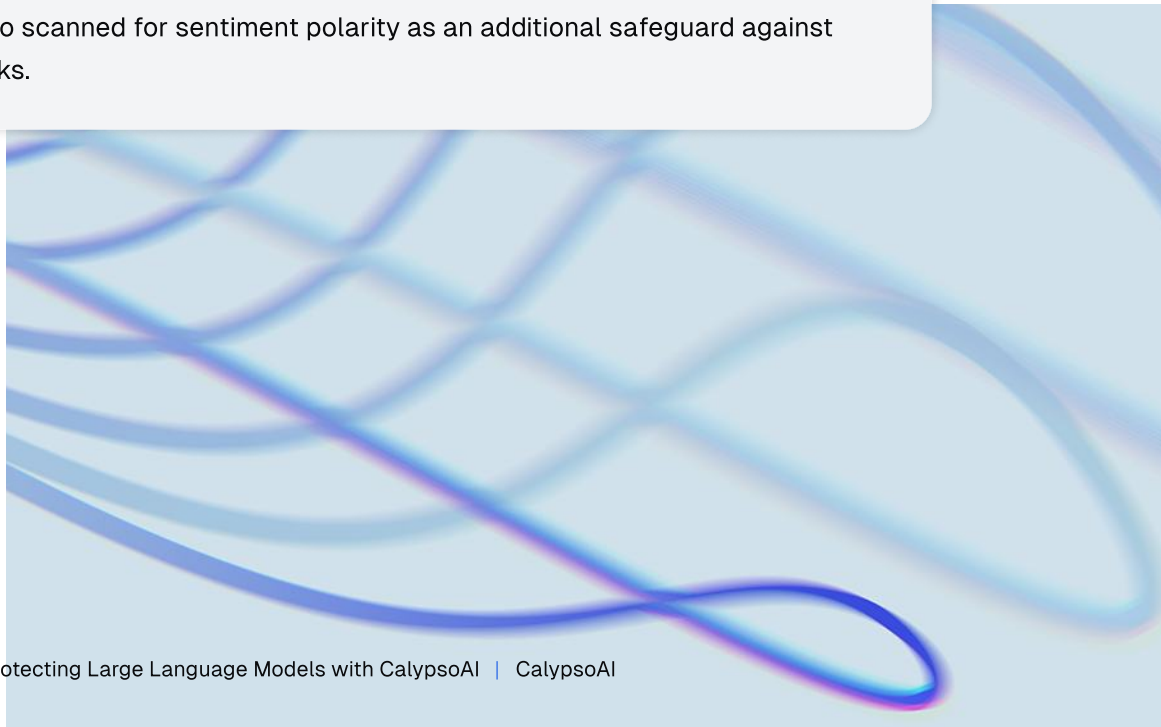
The Threat:

Prompt injections occur when threat actors use manipulative language structures—such as role-playing, reverse psychology, hypothetical scenarios, or extensive world-building—to exploit the model. These tactics can bypass control mechanisms, forcing the model to act outside its intended parameters. Thwarting prompt injections is essential to preserve the integrity and reliability of LLM applications.

How CalypsoAI Addresses This Threat

Preventing System Compromise

CalypsoAI scans every prompt and response for manipulative language structures. Our prompt injection scanner is continually updated to prevent threat actors from exploiting the model. Prompts are also scanned for sentiment polarity as an additional safeguard against sentiment-driven attacks.

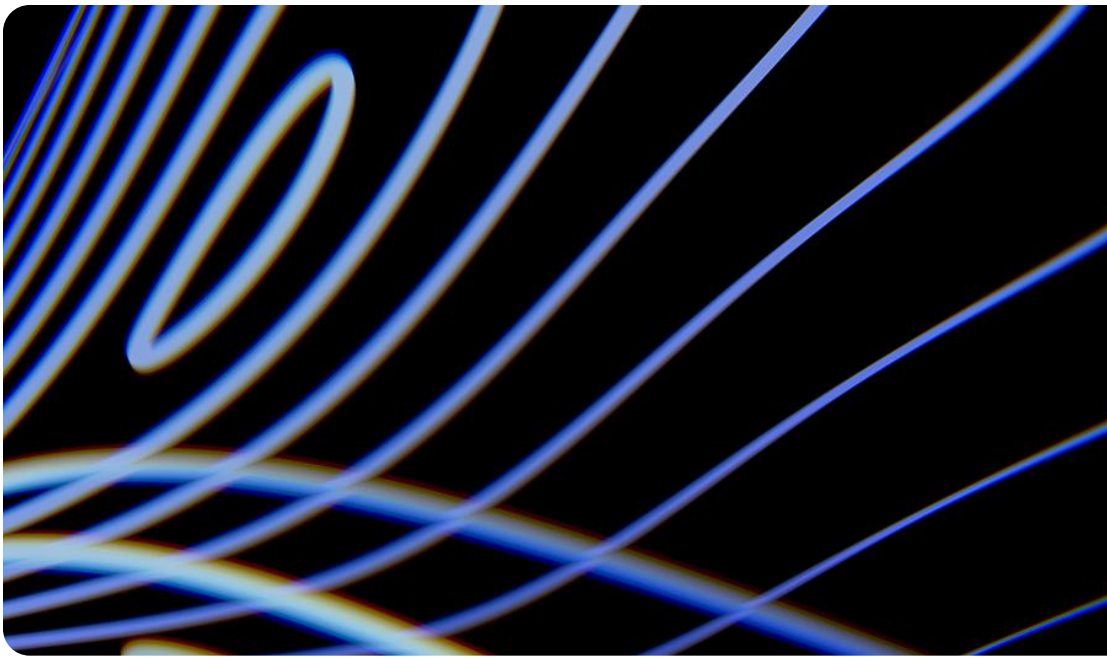


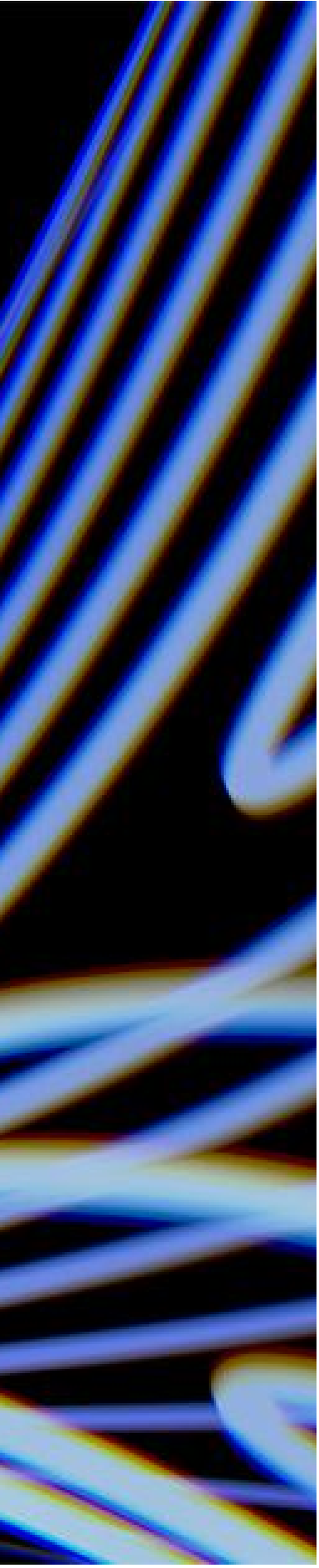
LLM02 - Insecure Output Handling

The Threat:

Insecure output handling arises when LLM-generated responses are accepted without adequate scrutiny, potentially exposing back-end systems to cyber threats. This vulnerability can lead to serious consequences, such as Cross-Site Scripting (XSS), Cross-Site Request Forgery (CSRF), Server-Side Request Forgery (SSRF), privilege escalation, or even remote code execution.

Insecure output handling that results in a user triggering malicious instructions can lead to the dissemination of harmful or inappropriate content or leakage of sensitive information. Implementing robust output filtering and validation mechanisms can mitigate these risks.





How CalypsoAI Addresses This Threat

Scanning for Malicious Responses

The CalypsoAI solution employs a rigorous scanning process for all LLM responses, identifying potentially harmful import statements and paying close attention to those containing code.

CalypsoAI's Source Code scanners analyze both the prompt and response and support a wide range of programming languages, including Python, Java, JavaScript, PHP, SQL, TypeScript, Go, C, C++, C#, CSS, HTML, VBA, and Bash. These scanners assess code structures and commands, identifying potential malicious or harmful components. Any response found to contain such elements is prevented from entering the system.

By blocking potentially harmful responses, CalypsoAI protects the application from security threats, including those resulting from insecure output handling.

Employing Human Verification

CalypsoAI offers a customizable Human Verification feature, allowing administrators to mandate manual review of all content in a response. This adds an extra layer of security, ensuring that even the most subtle anomalies are detected and addressed. The human verification process acts as a critical checkpoint, especially in scenarios in which automated systems might not fully capture the complexity or context of the output.

LLM03 - Training Data Poisoning

The Threat:

Training data poisoning is a critical vulnerability in LLM applications, occurring when the training data, sourced from repositories like Common Crawl, WebText, and OpenWebText, is manipulated. This tampering can introduce vulnerabilities or biases, thereby compromising the security, effectiveness, or ethical behavior of the LLM.

Poisoned data can skew the model's outputs, leading to flawed or biased decisions.

How CalypsoAI Addresses This Threat

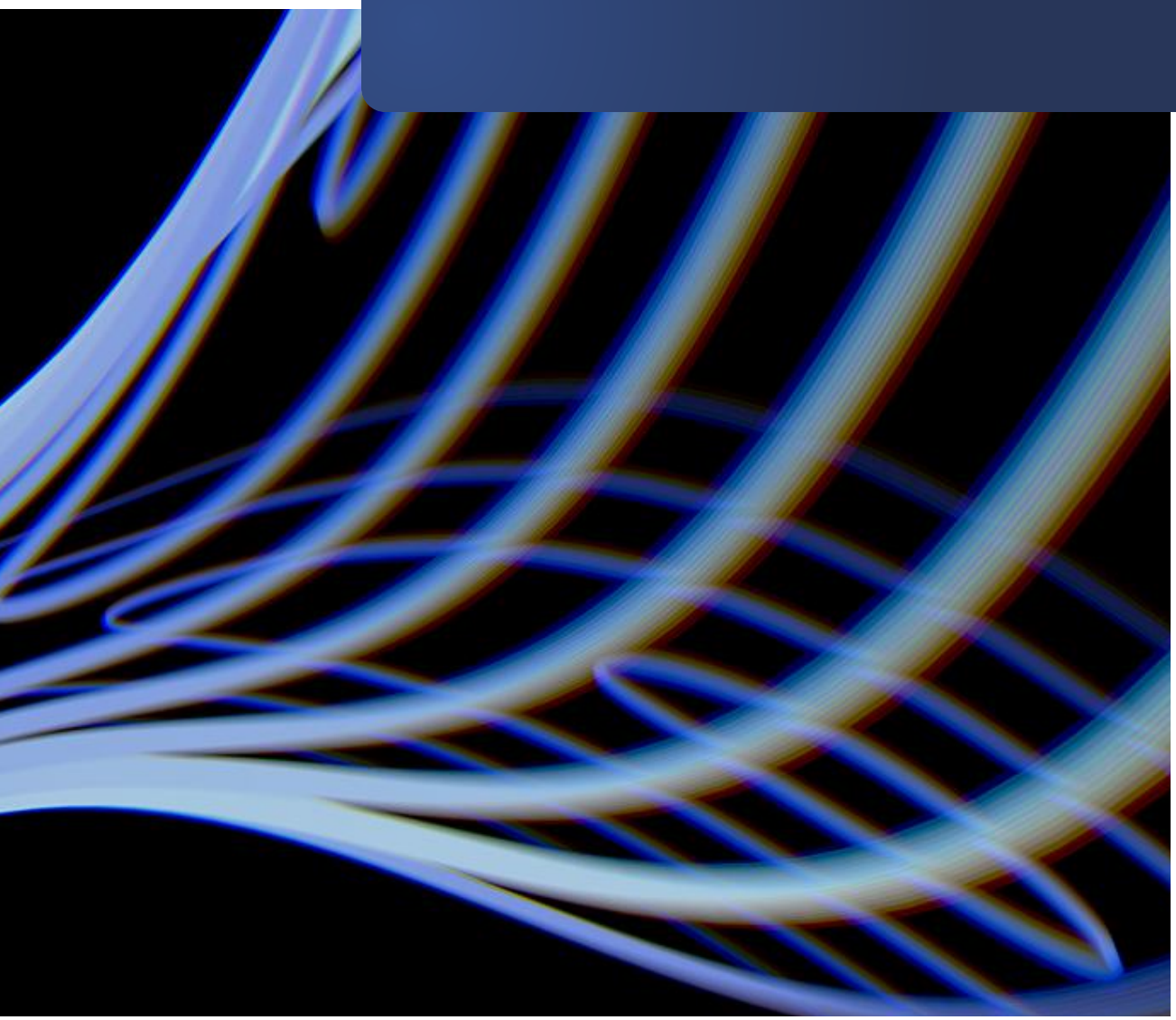
Integrating an Additional Layer of Review

By integrating an additional layer of review and validation, CalypsoAI's Human Verification feature plays a pivotal role in ensuring the accuracy and reliability of LLM responses. This feature operates on a dual-layer review process, in which responses are subjected to review by human experts, which is critical in identifying and rectifying any biases or anomalies that might be overlooked by automated systems alone.

LLM04 - Model Denial of Service

The Threat:

Model Denial of Service (DoS) attacks are a growing concern for LLM administrators; attackers can initiate resource-heavy operations that lead to service degradation or increased costs. LLMs' resource-intensive nature and the unpredictability of user inputs leave these systems particularly vulnerable to such disruptions.



How CalypsoAI Addresses This Threat

Providing Layers of Control

CalypsoAI offers enhanced auditability features, enabling administrators to monitor system usage effectively. This insight allows for the identification of irregular patterns that could indicate potential threats, which is crucial for the early detection of DoS attacks.

Strategizing Resource Allocation and Model Rotation

Whether models face intentional attacks or just very high demand, CalypsoAI mitigates the risk of DoS attacks by enabling administrators to strategize resource allocation and implement LLM rotation, allowing them to intelligently route traffic based on model load. This approach ensures that no single model bears excessive load, thus maintaining operational efficiency and preventing system overuse.

By optimizing resource distribution and enabling model rotation, CalypsoAI can enhance the resilience and reliability of LLM applications.

LLM05 - Supply Chain Vulnerabilities

The Threat:

Supply chain vulnerabilities in LLM applications arise from integrating potentially vulnerable components or services. Incorporating third-party datasets, pre-trained models, and plugins can introduce security risks. These vulnerabilities can compromise the entire application lifecycle, leading to potentially serious security breaches.

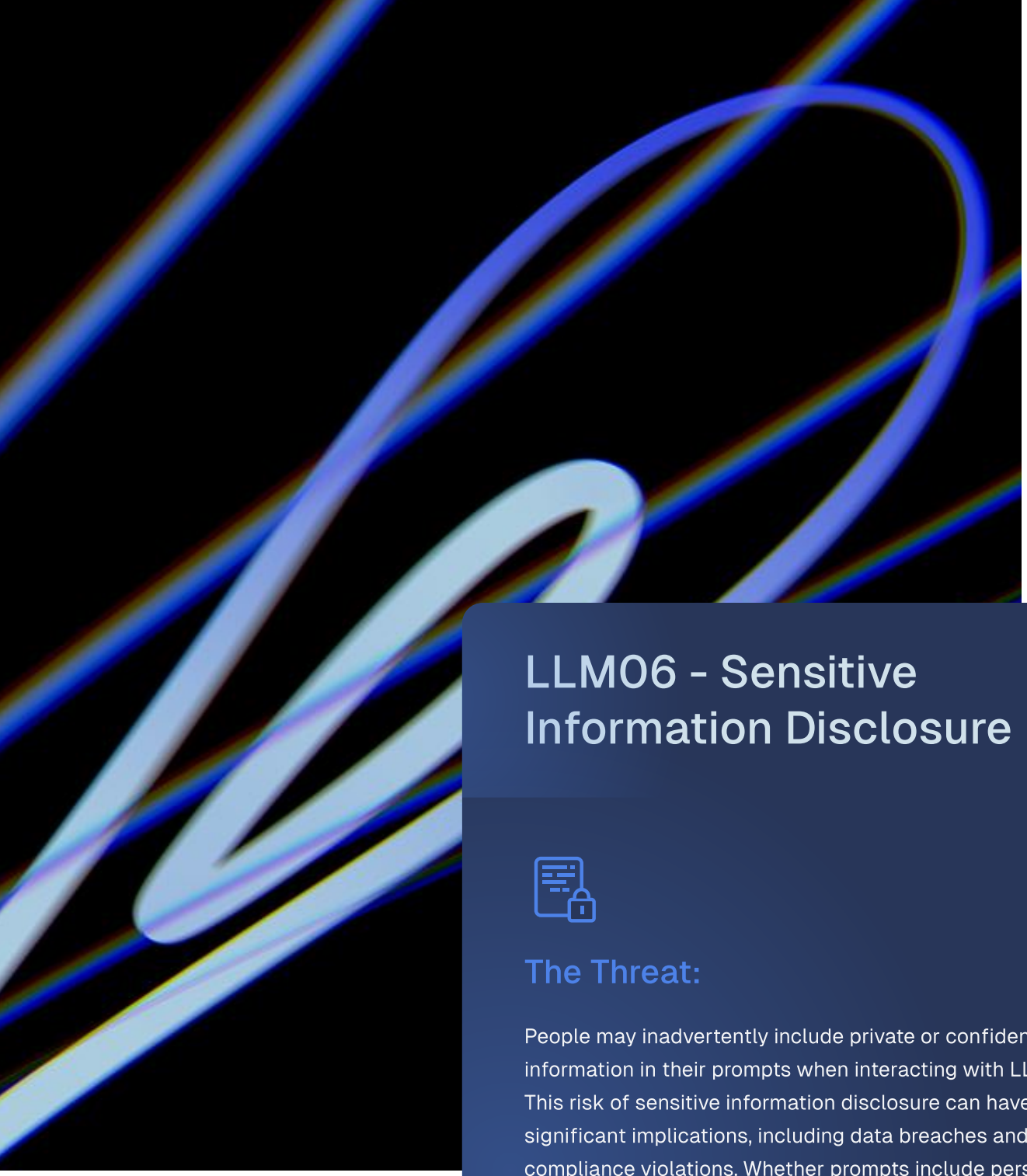
How CalypsoAI Addresses This Threat

Building a Reliable Information Taxonomy

CalypsoAI enhances security by allowing organizations to customize LLM usage. It achieves this by allowing specific routing rules to be set for different groups or teams based on their trust levels and needs. This capability is instrumental in constructing and maintaining a trusted information taxonomy within the organization. CalypsoAI can enhance data integrity by regulating access and usage according to policy-based access management control levels that can be integrated with OAuth and Microsoft Active Directory, ensuring that sensitive or critical information is handled appropriately.

Preventing Unauthorized Data Usage

CalypsoAI ensures only verified and trusted information is utilized within the LLM processes. This preventive measure shields the system from potential threats arising from compromised third-party components. CalypsoAI can play a vital role in safeguarding the LLM against vulnerabilities by preventing unauthorized or malicious data from affecting the supply chain.



LLM06 - Sensitive Information Disclosure

The Threat:

People may inadvertently include private or confidential information in their prompts when interacting with LLMs. This risk of sensitive information disclosure can have significant implications, including data breaches and compliance violations. Whether prompts include personal details, proprietary data, or other sensitive information, such accidental disclosures pose a serious challenge in maintaining the confidentiality and integrity of communications and data processing within LLM environments.

How CalypsoAI Addresses This Threat

Utilizing a Diverse Range of Scanners

To address the risk of sensitive information disclosure, CalypsoAI deploys advanced scanning algorithms that identify and flag potential instances of sensitive data within user inputs. CalypsoAI then prevents this information from being sent to the LLM.

CalypsoAI's scanners include:

1. Personally Identifiable Information (PII) Scanner: Detects and flags PII within prompts to prevent unintended disclosure.
2. Named Entity Scanner: Detects names of entities, such as specific people, organizations, and locations, preventing unauthorized disclosure of sensitive information regarding the organization's relationship with the entity.
3. Secrets Scanner: Identifies and blocks exposure of sensitive information like passwords or encryption keys, ensuring confidentiality.
4. Source Code Scanner: Secures proprietary or sensitive source code to prevent unauthorized public disclosure.
5. Data Loss Prevention (DLP) Scanner: Customizable to include company-specific terms, it prevents prompts containing proprietary and other company data from being sent to an LLM.
6. Prompt Injection Scanner: Detects and mitigates manipulation attempts in LLM prompts intended to circumvent internal safeguards regulating model behavior.

LLM07 - Insecure Plugin Design

The Threat:

LLM plugins often lack sufficient access control mechanisms and can have insecure inputs. This deficiency in application control can make plugins susceptible to exploitation, potentially leading to severe consequences, such as remote code execution. Ensuring these plugins' security is crucial to maintaining the overall integrity of LLM systems.

How CalypsoAI Addresses This Threat

Restricting Plugin Access

CalypsoAI employs stringent access controls for users and groups. By actively managing who can use certain models, the tool significantly reduces the risk of malicious exploitation of LLM plugins. This approach ensures that only authorized personnel have access to critical functionalities, safeguarding the system from potential misuse or attacks.

Scanning Plugin Responses for Malicious Content

CalypsoAI rigorously scans responses from plugins to detect any malicious content, including harmful commands and code. By identifying and removing these potential threats, CalypsoAI effectively safeguards the system against vulnerabilities that might arise from plugin interactions, maintaining the reliability and security of the LLM applications.

Continuously Updating Language Recognition

CalypsoAI regularly updates its suite of recognized computer languages to stay ahead of evolving threats. This continual enhancement improves the tool's ability to detect and neutralize potential threats in plugin responses, ensuring its defense mechanisms remain effective against the latest cyber risks.

Preventing Unauthorized Code Integration

CalypsoAI actively safeguards against integrating unauthorized or malicious code from plugins into the LLM or the hosting system. This essential function ensures that any potentially harmful code is blocked before it can compromise the system, upholding the overall safety and reliability of the LLM applications.

LLM08 - Excessive Agency

The Threat:

Excessive agency in LLM-based systems can lead to unintended consequences, often stemming from granting too much functionality, access, or autonomy to these systems. This issue poses a significant challenge in ensuring that LLMs operate within their intended parameters and do not inadvertently cause harm or security breaches.

How CalypsoAI Addresses This Threat

Implementing Policy-Based Access Controls

Implementing policy-based access controls is a crucial means of addressing this. CalypsoAI enables fine-grained permission settings, allowing organizations to define and enforce who can access specific functionalities based on roles and responsibilities.

This feature minimizes the risk of unauthorized access and streamlines the management of LLM systems, ensuring that each user has the appropriate level of access needed for their role.

> Unrestricted access can lead to potential security risks and management challenges in LLM-based systems.

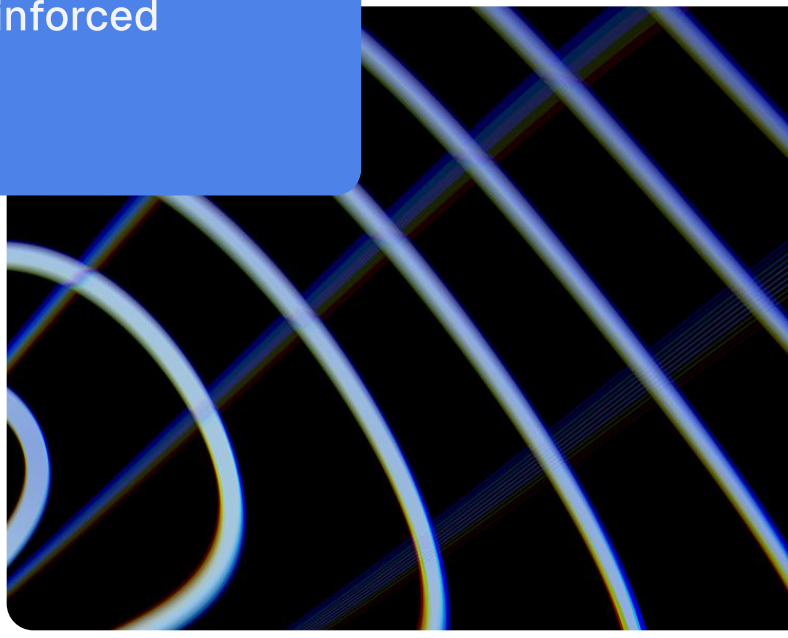
Integrating with Rate Limits and Access Safeguards

CalypsoAI enables administrators to use rate limits and other access-related safeguards to effectively restrict and guide user behavior. These integrations are key to managing the frequency and type of interactions users can have with the LLM, preventing misuse, and reducing the likelihood of unintentionally harmful actions.

Providing Detailed User Interaction Insights

Having a broad view of user interactions can be highly beneficial for organizations using LLMs. CalypsoAI enables this insight by retaining each query and response, providing a comprehensive overview of user interactions. Each interaction is scored against scanner thresholds, which aids in monitoring and analyzing the nature and impact of these exchanges. When reviewed across the organization, such information aids in identifying patterns, assessing system effectiveness, and ensuring compliance with company policies or standards.

Organizations can fine-tune their LLM applications based on actual usage data, leading to improved decision-making, enhanced user experiences, and reinforced system security.



LLM09 - Overreliance

The Threat:

An overreliance on LLMs without a counterbalance of appropriate oversight can lead to organizational challenges, including the spread of misinformation, miscommunications, legal issues, and security vulnerabilities, due to incorrect or inappropriate content generated by the models.

How CalypsoAI Addresses This Threat

Reviewing Model-Generated Code

CalypsoAI's Malware Source Code scanner reviews code generated by models. This scanner includes a broad array of continually updated computer languages, ensuring comprehensive coverage and up-to-date analysis. It rigorously examines the syntax and potential dangers of model-generated code, identifying any elements that could pose risks to the system.

Enforcing Disclosure Agreements

CalypsoAI allows organizations to enforce specific rules for LLM usage through customizable disclosure agreements to ensure users interact with LLMs responsibly and in compliance with organizational policies and legal standards.

Detecting Toxic and Banned Content

CalypsoAI identifies and blocks any outgoing or incoming content that might be considered harmful, offensive, or in violation of company policy.

LLM10 - Model Theft

The Threat:

Model theft is a critical concern for organizations deploying fine-tuned LLMs and can include breaches such as unauthorized access to, copying, or exfiltration of proprietary models. This form of intellectual property loss violates privacy and security protocols and poses significant risks to the organization's competitive advantage and operational integrity. The theft of these advanced models can lead to substantial economic losses, as the proprietary technology and research behind the models represent significant investments in time, resources, and expertise.

How CalypsoAI Addresses This Threat

Enabling Real-Time Monitoring and Implementing Rate Limiting

CalypsoAI provides administrators the ability to monitor, audit, and manage LLM interactions in real time, enhancing control and visibility. CalypsoAI also allows administrators to implement rate limits for the number of calls to a model. By restricting the frequency of model access, the tool effectively reduces the risk of mass data extraction and model theft, and helps maintain system performance and efficiency.

Differentiating Between Rate Limiting and Prompt Injections

It's important to note that rate limiting and prompt injections are distinct concerns, each with unique implications for model security.

1. Rate limiting controls the quantity of model interactions and prevents excessive model querying, which could be a sign of an attempt to copy or exfiltrate the model.

2. Prompt injection scanning detects and mitigates attempts to manipulate model responses, preventing the model from generating responses that could inadvertently reveal sensitive information about or proprietary aspects of the model itself.

Defend Against the OWASP Top 10 for LLM Applications with CalypsoAI

In the rapidly evolving landscape of cybersecurity, staying ahead of threats is paramount. CalypsoAI emerges as a comprehensive solution, uniquely tailored to address the vulnerabilities highlighted in the OWASP Top 10 for Large Language Model Applications. With its suite of specialized scanners—from PII detection to prompt injection mitigation—and fine-grained controls, CalypsoAI is an unparalleled defense mechanism for LLM applications.

Take action now to fortify your LLM applications against emerging cyber threats.

Book a demo today to see firsthand how CalypsoAI can transform your approach to LLM security.

CALYPSOAI