CALYPSOAI

Sharon Ober | Regional Manager | 703-851-6938 | sharon.ober@calypsoai.com

Generative Artificial Intelligence (AI), specifically Large Language Models (LLMs), have gained significant prominence across various industries, owing to their capacity to deliver valuable insights, automate tasks, and enhance efficiency. OpenAI's ChatGPT reached the 100 million user mark in January 2023, merely three months after its launch in November 2022. This rate of adoption has been significantly faster than that of Facebook or Twitter. Although there is still limited data regarding the use of LLMs within organizations (both public and private), the increasing number of articles, conferences and private events discussing their implementation in corporate environments or within government entities suggests a similarly widespread adoption.

However, this logarithmic expansion has also introduced new challenges and risks for corporate organizations, such as exposing company Intellectual Property (IP), generating inappropriate or harmful content, and propagating misinformation, as well as the need for monitoring and moderating user interactions, among other concerns. LLMs also seem to be vulnerable to specific types of attacks aiming at degrading the user experience, infiltrating the organization's IT infrastructure, or propagating misinformation or disinformation.

Future Tense, LLC (DBA CalypsoAI) has spent the last 4+ years researching and developing novel AI/ML Testing, Evaluation, Validation and Verification (TEVV) and security solutions, many of which are focused on adversarial AI. We build upon academic research, developing and productizing novel AI test capabilities, which have undergone rigorous testing and provide customers with 60x-80x performance gains over conventional manual processes for model TEVV. We have provided TEVV solutions for various data types including, but not limited to, satellite imagery, X-ray imagery, MMW, IR, FMV, CV, LLMs, and SAR. The company currently holds two USPTO approved patents: 1) US 10,846,407 B1: Machine Learning Model Robustness Characterization and 2) US 10,839,268 B1: Artificial Intelligence Adversarial Vulnerability Audit Tooling.

In addition to deep data science and research expertise, CalypsoAI has several innovative commercial technical capabilities that we leverage in RDT&E and operational engagements.

1. Specific to LLMs, CalypsoAI has developed Moderator™, our ground-breaking tool provides a full wraparound security perimeter, protecting your organization from known and emerging LLM risks. It enables observability with full engagement tracking and auditability for both individual users and user groups, providing insights into usage, resource consumption, and content. Multi-layered customizable scanners ensure organizational policies and standards are applied to all user prompts and LLM responses. This capability solves four key LLM problems: (1) unmoderated user interactions, harmful content, and misinformation; (2) IP protection and data privacy; (3) security vulnerabilities; and (4) tracking and analyzing user usage.

2. VESPR Validate™, an automated and extensible testing platform, ensures AI/ML can securely achieve mission goals in real-world conditions. It stress tests models against real world performance (i.e. weather conditions, blur), privacy (model vulnerability to leaking sensitive training data), and adversarial security (white box and black box, including model inversion and evasion tests). We provide a generalized API-integration portal to connect models from any environment and then supply a battery of tests built on a microservices architecture. Capabilities can be smoothly integrated with a range of third party solutions and across multiple deployment environments. R&D is underway to add model runtime monitoring capabilities for adversarial attacks; performance and operational metrics; and identification of framework errors.