CyberFlood AI Inference Testing

Validate the performance and security of AI Inference Infrastructures and LLM Applications

Generative AI is transforming industries, driving applications from conversational assistants to advanced analytics. As organizations scale AI workloads, they face pressure to deliver low-latency, high-throughput inference while ensuring security and cost efficiency. Performance issues or downtime can lead to financial loss and erode trust.

Security risks are rising, with breaches exposing sensitive data from AI models, underscoring the need for robust governance and testing.

Traditional methods can't keep pace: inference workloads are dynamic, resource-intensive, and sensitive to network and security configurations. Without rigorous validation, organizations risk latency spikes, resource mis-allocation, and attacks like prompt injection or denial-of-service.

Solutions Overview

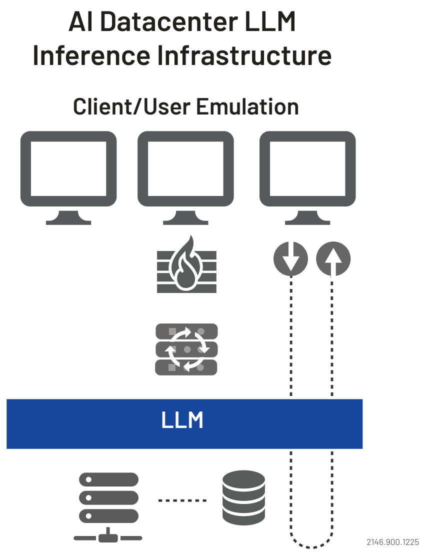
CyberFlood delivers scalable, high-fidelity client emulation to generate realistic AI inference workloads. It simulates real user interactions with LLMs, LLM APIs, and backend inference services using advanced stateful session modeling. Each client profile offers extensive configurability, including:

- Customizable LLMs and API endpoints
- Authentication options
- Dynamic prompt generation
- Configurable conversation depth and multi-turn interactions

This enables precise, repeatable testing of AI inference performance under real-world conditions.

CyberFlood's dynamic load generation and advanced load profile definition enable test engineers to model massive concurrencies and bursty inference traffic, emulating realistic production-scale user and application behavior. Combined with the flexibility to define extensive lists of prompts of varying length, conversational turns, and keywords, this ensures the testing reveals memory pressure thresholds, GPU saturation, and inference throughput degradation points, as well as basic response accuracy via keyword match.

The flexible prompt list also enables users to emulate negative workloads with adversarial and malicious prompts, and, when paired with CyberFlood stateful DDoS scenarios, to validate the security controls and policies in place to safeguard the inference infrastructure and LLM applications.



CyberFlood AI Inference Testing

CyberFlood's purpose-built AI Inference testing capabilities enable LLM vendors, cloud AI infrastructure providers, and organizations implementing AI applications to test Generative AI inference systems under real-world conditions. It enables teams to validate scalability, performance, and security, ensuring that AI-powered applications are production-ready and resilient.

CyberFlood AI Inference testing enables users to:

Validate the End-to-End AI Inference Infrastructure

Evaluate how network components, API gateways, firewalls, ADCs, GPU compute capacity and security controls impact LLM inference throughput, latency, and accuracy.

Benchmark LLM Performance Under Real-World Load Conditions

Test context length limits, prompt/response size boundaries, multi-model serving, tokenizer overhead, pre-prompt token calculation, and long-running conversational sessions.

Ensure the Security of LLM Applications

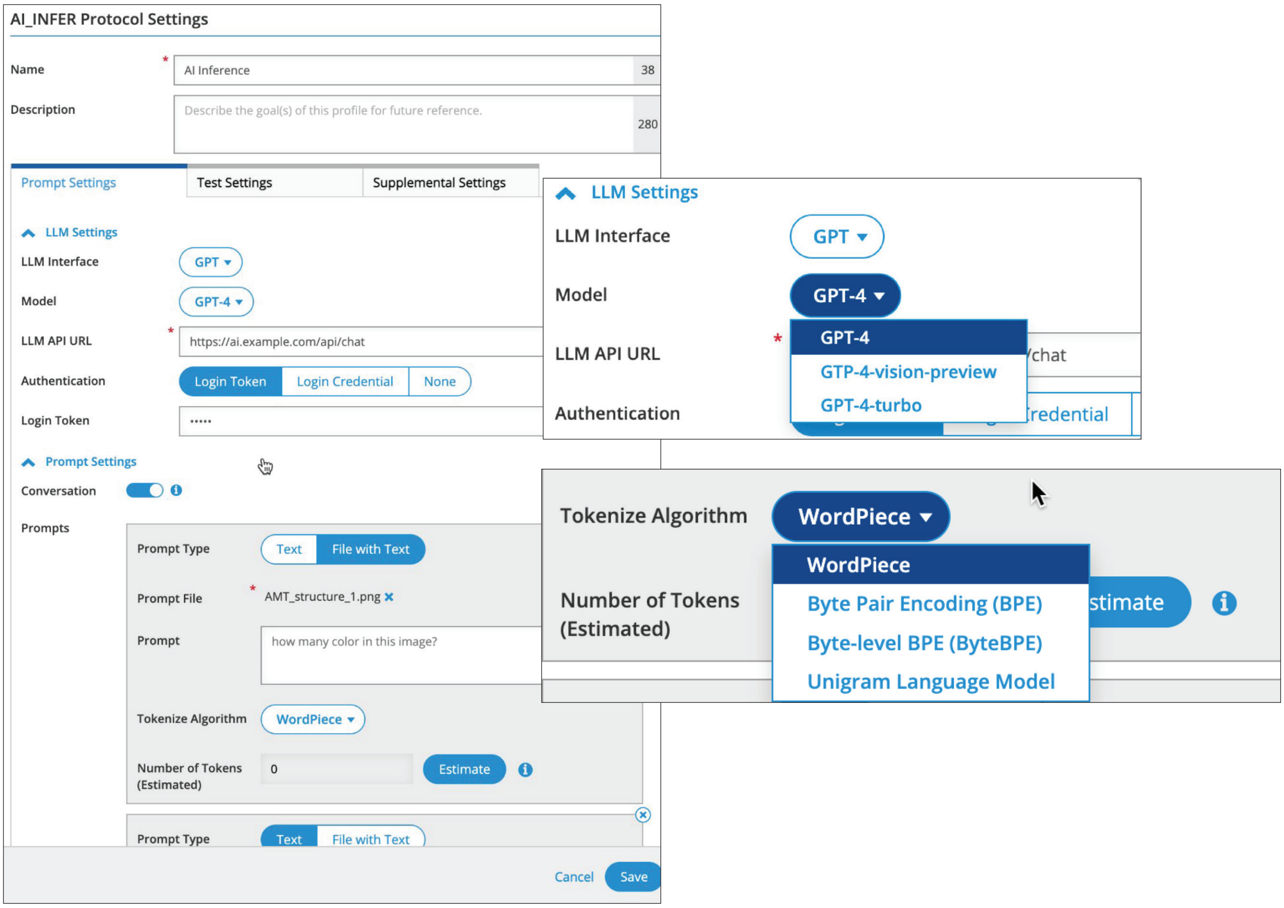
Emulate prompt injection attempts, malformed payloads, high-rate API abuse, and other cyber threat scenarios.

Right-size Scalability and Resilience

Characterize inference cluster behavior under concurrency spikes and diverse prompt profiles to expose KV-cache memory pressure, GPU/accelerator saturation, batching limits, and scheduling bottlenecks.

Key Capabilities and Benefits

- Stateful One-arm clients emulate real user interactions with LLMs, APIs, and backend inference services using stateful session modeling.
- Model multi-step user journeys using both user-defined and pre-built prompt lists to orchestrate complex query chains, multi-turn conversations, tool calls, and agentic workflows.
- With varying prompts lists to push tokens, context windows, and prompt-length limits while injecting adversarial, abusive, or malformed prompts to validate guardrails.
- Simulate massive concurrency and bursty inference traffic to emulate production-scale application behavior
- Real-time metrics to measure input and output tokens per second, TTFT, end-to-end latency, throughput, concurrency, bandwidth, and basic response-accuracy scoring with automated response parsing.
- Comprehensive REST API support for automation and CI/CD integration.



AI_INFER Protocol Settings

Name * AI Inference 38

Description: Describe the goal(s) of this profile for future reference. 280

Prompt Settings | Test Settings | Supplemental Settings

LLM Settings

LLM Interface: GPT

Model: GPT-4

LLM API URL * https://ai.example.com/api/chat

Authentication: Login Token | Login Credential | None

Login Token: •••••

Prompt Settings

Conversation

Prompts

Prompt Type: Text | File with Text

Prompt File * AMT_structure_1.png

Prompt: how many color in this image?

Tokenize Algorithm: WordPiece

Number of Tokens (Estimated): 0

Prompt Type: Text | File with Text

CyberFlood Advanced Inference Configuration

Technical Specifications

LLM Models	OpenAI/gpt-4.1, Ollama/llama
LLM Profile Settings	TLS Encryption options
	Authentication
	Prompt Source (user input or from file)
	Conversation Mode
	Tokenize Algorithm
Prompt Setting	Text, Image, Keyword Match, Token Estimations
Load Types	Simulated Users
Key Metrics	Total Prompts Token
	Total Prompts Sent
	Total Response Token
	Load Duration
	Prompt Token Duration
	Prompts Executed
	Prompt Token Per Second
	Prompts Per Second
	Response Tokens Per Second
	Response Token Usage
	Time to First Response Token with 95/96/99 Percentiles
	Time to Last Response Token with 95/96/99 Percentiles
Supported Test Devices	CyberFlood Virtual
	CF400 and CF30 planned

Contact Us: +1 844 GO VIAVI | (+1 844 468 4284). To reach the VIAVI office nearest you, visit viavisolutions.com/contact

viavisolutions.com