CyberFlood AI Inference Testing

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

AI-powered applications from conversational assistants and copilots to autonomous agents are becoming the primary interface between organizations and their customers. As organizations scale AI workloads, the network and compute infrastructure must deliver low-latency, high-throughput inference while resisting an entirely new class of security threats. Performance failures or security breaches in production erode user trust, create financial liability, and undermine the AI investment.

Traditional application testing methods can't keep pace. AI inference workloads are fundamentally different from conventional web traffic, they are bursty, dynamic, resource-intensive, and highly sensitive to network topology, security configurations and GPU/memory contention. Without rigorous purpose-built validation, organizations risk latency spikes under concurrency, resource mis-allocation, and exposure to attacks like prompt injection or denial-of-service that target the inference pipeline itself.

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 with support for OpenAI
- TLS encryption and flexible authentication options
- Dynamic prompt generation from user-defined or file-based prompt
- Configurable conversation depth and multi-turn interactions
- Text, image-based, and audio/video prompts for multi-modal inference testing

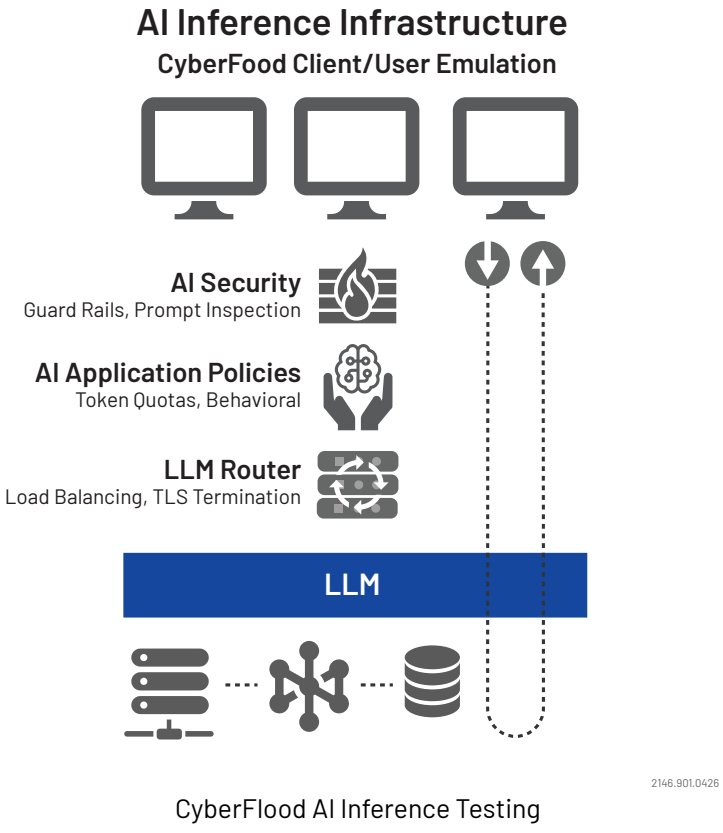
This enables precise, repeatable testing of AI inference performance under real-world conditions.

CyberFlood's dynamic load generation and advanced load profile definition enables AIOps and Quality Assurance 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 Inference testing reveals memory pressure thresholds, GPU saturation, and inference throughput degradation points, as well as basic response accuracy via keyword matching and pattern validation.

CyberFlood's purpose-built AI Inference testing capabilities enable LLM vendors, Neo-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.

End-to-end Inference Infrastructure Validation

Before a user prompt reaches the LLM, it traverses multiple infrastructure layers — each of which introduces its own potential failure modes, latency contributions, and security enforcement points. Comprehensive testing must validate the entire chain under load, not just the model endpoint in isolation.



CyberFlood AI Inference Testing

CyberFlood generates highly scalable inference traffic that flows through the complete infrastructure path, from AI security gateways and guardrails, through the AI application policies that are rate-limiting token usage and identifying non-human behavioral patterns, past the LLM Router layer to dynamically route prompts to gain cost efficiency and performance, and finally into the AI inference servers and GPU clusters via the LLM. This full-loop approach exposes how each component affects throughput, latency, and response accuracy under production-scale conditions.

The Expanded AI Security Attack Surface

Front-end AI applications introduce entirely new attack vectors that conventional security testing does not address. AI inference infrastructure must be validated against threats specific to the LLM serving pipeline:

- **Prompt injection attacks** — Crafted inputs that embed hidden instructions to manipulate model behavior, ex-filtrate data, or bypass safety controls.
- **Context window overflow** — Deliberately oversized inputs designed to exhaust model context limits, causing degraded outputs, crashes, or code injection.
- **High-rate API abuse and DDoS** — Volumetric attacks targeting inference endpoints to saturate GPU compute, exhaust memory, and degrade service for legitimate users.
- **Adversarial and malformed payloads** — Inputs crafted to exploit tokenization edge cases, bypass guardrails tuned for common languages, or trigger unintended model behaviors.
- **Data exfiltration via model manipulation** — Multi-turn conversation strategies designed to extract training data, PII, or proprietary information from the model's responses.

Validating AI security requires emulating these attack patterns at production scale, simultaneously with legitimate workloads, to verify that security controls maintain efficacy without introducing unacceptable performance degradation or false-positive blocking of legitimate requests.

CyberFlood uniquely combines AI inference performance testing with proven stateful DDoS and cyber threat emulation capabilities. This integrated approach enables teams to validate security policies, guardrails, and infrastructure resilience under combined legitimate and adversarial traffic — a critical requirement that performance-only testing tools cannot deliver.

Right-Sizing Inference Infrastructure Investments

AI inference is a recurring operational expenditure, industry analysis suggests that inference consumes 80–90% of total AI compute spend over a model's production lifecycle. Every over-provisioned GPU, sub-optimal batching configuration, and mis-sized KV-cache allocation wastes budget. Rigorous testing under realistic load conditions reveals where resources are genuinely needed and where cost savings can be captured, enabling data-driven infrastructure decisions backed by actual performance data, not synthetic benchmarks.

CyberFlood AI Inference Key Use Cases

Validate the End-to-End AI Inference Infrastructure

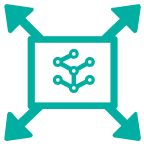
Evaluate how network components, API gateways, firewalls, ADCs, GPU compute capacity and security controls impact LLM inference throughput, latency, and accuracy.

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.

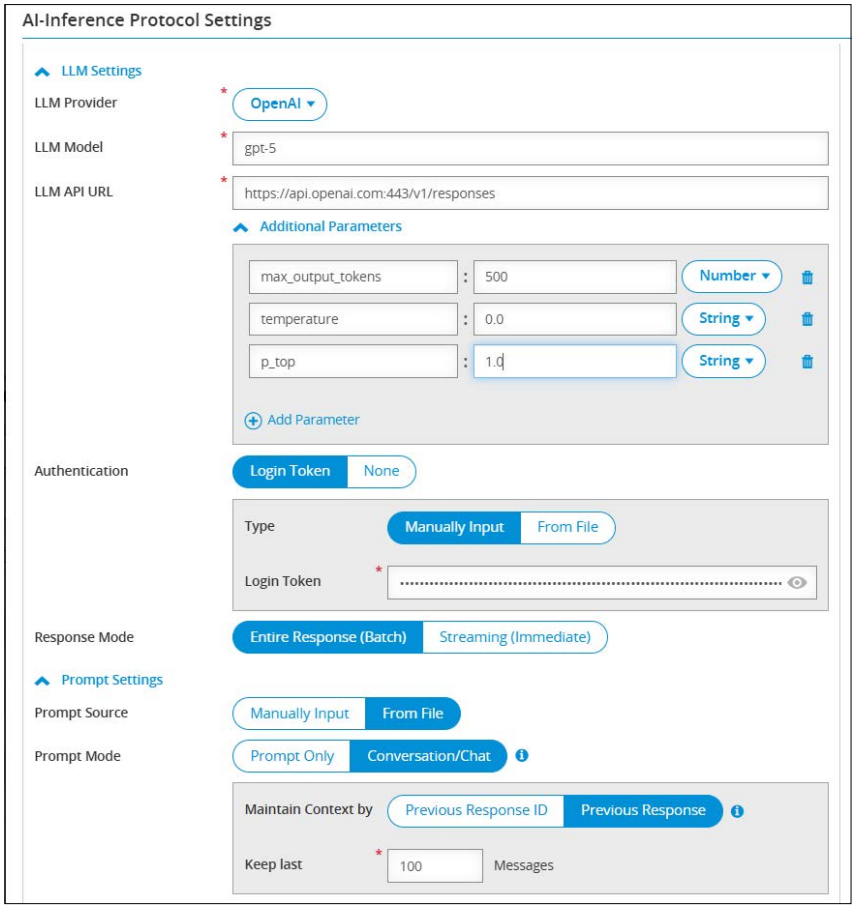


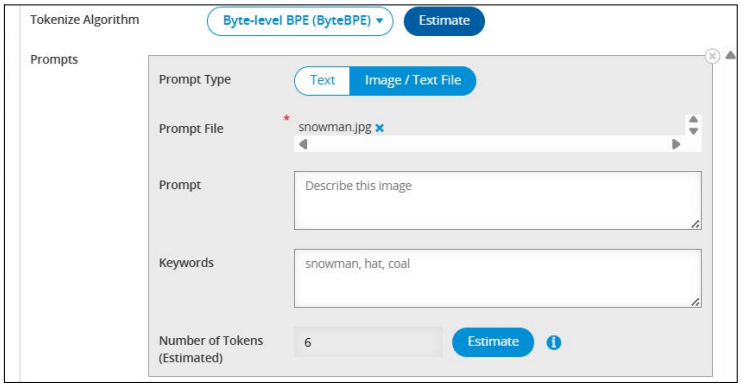
Right-size Scalability and Resilience

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 client** — Emulate real user interactions with LLMs, APIs, and backend inference services using stateful session modeling that maintains conversation context across multi-turn interactions. Programmable user journeys — 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.
- **Boundary and adversarial testing** — Use customizable prompt lists to push tokens, context windows, and prompt-length limits while injecting adversarial, abusive, or malformed prompts to validate guardrails and security controls.
- **Production-scale load emulation** — Simulate massive concurrency and bursty inference traffic with advanced load profile definition including custom phases (ramp, burst, steady state) to emulate production-scale application behavior
- **Multi-modal prompt support** — Test with text-based, image-based and video/audio-based input prompts to validate multi-modal inference endpoints and vision-language models.
- **Real-time inference-native metrics** — Measure input and output tokens per second, Time to Frist Response Token (TTFT), Time Per Output Token, Time to Last Response Token (TTLT), token usage throughput, concurrency, bandwidth, and basic response-accuracy scoring with automated keyword matching and response parsing.
- **CI/CD integration** — Comprehensive REST API support for test automation enabling integration into continuous validation pipelines and modern DevOps workflows.





CyberFlood Advanced Inference Configuration

Technical Specifications

LLM Models / Servers	OpenAI, OpenAI compatible mode, Ollama
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. Custom prompt parameters using a name/value type allows the user to add parameters specific to the LLM under test, such as temperature, top_p, max tokens
Load Types	Simulated Users
	Prompt Token Usage, Total Token Used
	Prompts Executed
	Prompt Token Per Second
	Prompts Per Second
	Response Tokens Per Second
	Response Token Usage, Time Per Output Token
	Time to First Response Token with Long Tail 95/96/99 Percentiles
	Time to Last Response Token with Long Tail 95/96/99 Percentiles
Supported Test Devices	CyberFlood Virtual, CF30, CF400

Ordering Information

Part Number	Description
CF-SW-AI-INFERENCE	CyberFlood AI Inference Protocol for AMT
CF-SW-AI-INFERENCE-SUB	CyberFlood AI Inference Protocol for AMT Subscription

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

cyberflood-aiinterference-ds-hse-nse-ae
30194760 901 0426