CyberFlood AI 推理测试

验证 AI 推理基础设施和 LLM 应用程序的性能和安全性

生成式 AI 正在改变各个行业，推动从对话助手到高级分析等多种应用的发展。随着机构扩展 AI 工作负载，他们面临着在确保安全性和成本效益的同时提供低延迟、高吞吐量推理的压力。性能问题或宕机可能导致经济损失并削弱信任。

安全风险不断攀升，数据泄露事件频发导致 AI 模型中的敏感信息暴露，这凸显出强健治理与测试的迫切需求。

传统方法难以为继：推理工作负载具有动态性、资源密集性，且对网络和安全配置高度敏感。若缺乏严格验证，企业可能面临延迟激增、资源分配不当以及即时注入或拒绝服务等攻击风险。

解决方案概述

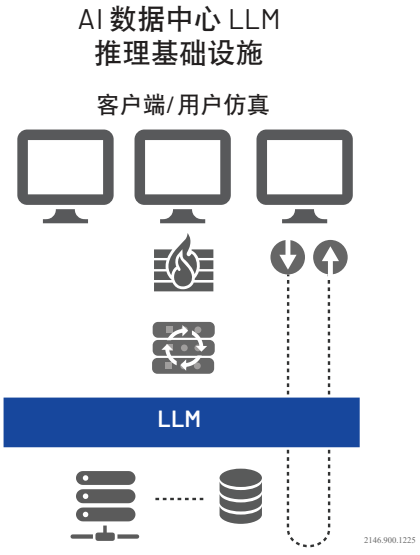
CyberFlood 提供可扩展的高保真客户端仿真功能，用于生成真实的 AI 推理工作负载。它通过先进的有状态会话建模技术，模拟真实用户与 LLMs、LLM API 及后端推理服务的交互过程。每个客户端配置文件都提供广泛的可配置性，包括：

- 可定制的 LLM 与 API 端点
- 身份验证选项
- 动态提示词生成
- 可配置的对话深度与多轮交互

这使得在真实条件下对 AI 推理性能进行精确、可重复的测试成为可能。

CyberFlood 的动态负载生成与先进负载配置文件定义功能使测试工程师能够模拟大规模并发及突发推理流量，真实再现生产级用户与应用行为。结合灵活定义不同长度提示词、对话轮次和关键词的广泛列表，该测试能够揭示内存压力阈值、GPU 饱和及推理吞吐量下降点，同时通过关键词匹配验证基本响应准确性。

灵活的提示词列表还支持用户通过对抗性和恶意提示词模拟负面工作负载，并结合 CyberFlood 有状态 DDoS 场景，验证现有安全控制措施与策略是否能有效保护推理基础设施及 LLM 应用程序的安全。



CyberFlood AI 推理测试

CyberFlood 专门构建的 AI 推理测试功能使 LLM 供应商、云 AI 基础设施提供商和实施 AI 应用的机构能够在真实世界条件下测试生成式 AI 推理系统。它使团队能够验证可扩展性、性能和安全性，确保 AI 驱动的应用程序可以投入生产并具有弹性。

CyberFlood AI 推理测试使用户能够：

验证端到端 AI 推理基础设施

评估网络组件、API 网关、防火墙、ADC、GPU 计算能力及安全控制措施对 LLM 推理吞吐量、延迟和准确性的影响。

真实负载条件下的基准 LLM 性能

测试上下文长度限制、提示词/响应响应长度区间、多模型服务、分词器开销、预提示词 Token 计算以及长时间运行的交互会话。

确保 LLM 应用程序的安全性

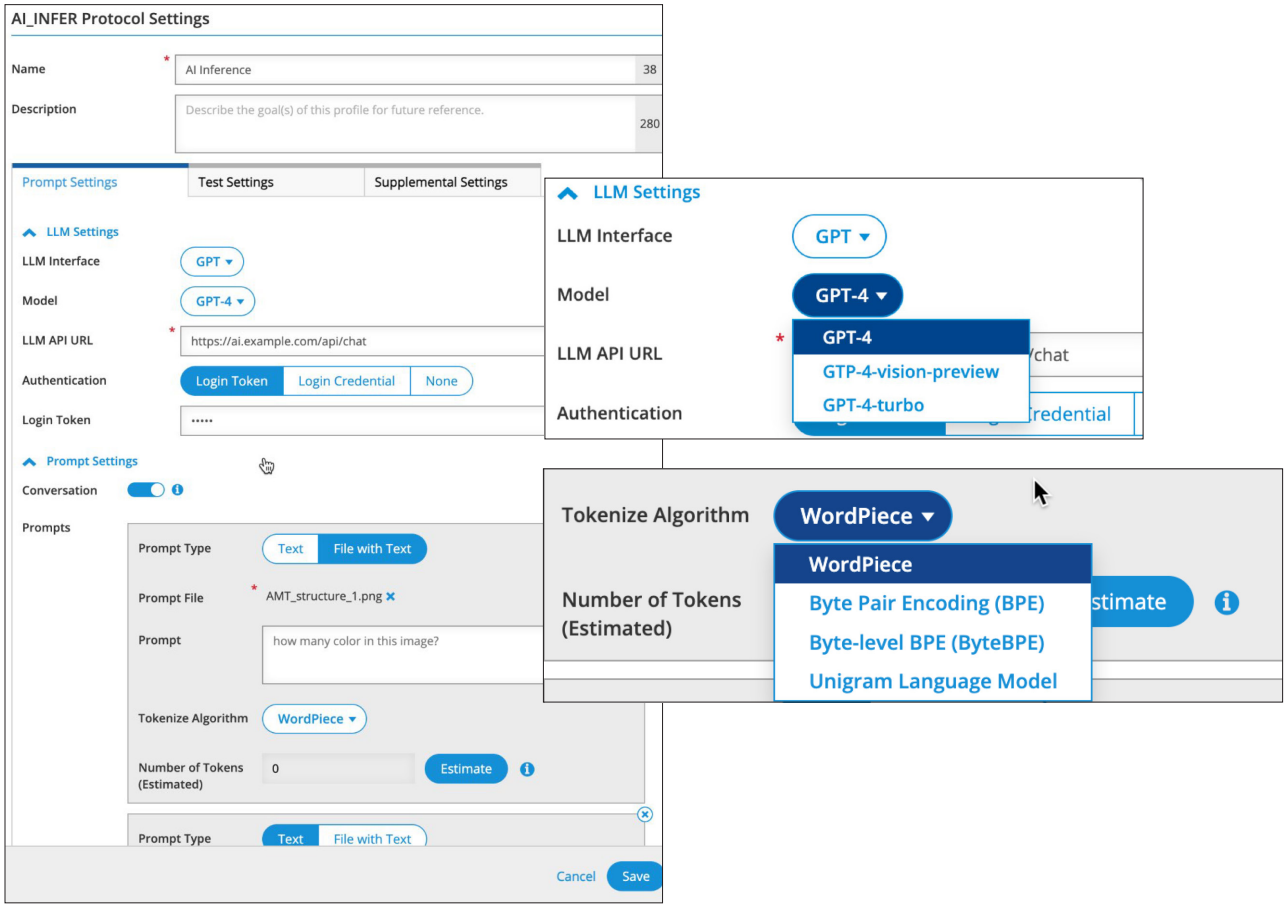
模拟提示词注入尝试、异常负载、高速率 API 滥用和其他网络威胁场景。

适当规模的可扩展性和弹性

分析并发激增和多样化提示场景下的推理集群行为，以揭示 KV 缓存内存压力、GPU/加速器饱和、批处理限制及调度瓶颈问题。

主要功能和优势

- 有状态单臂客户端使用有状态会话模型模拟真实用户与 LLM、API 和后端推理服务的交互。
- 利用用户自定义和预构建的提示词列表构建多步骤交互模型，以编排复杂的查询链、多轮对话、工具调用及代理工作流。
- 通过不同的提示列表来推送 Token、上下文窗口和提示长度限制，同时注入对抗性、侮辱性或格式错误的提示词以验证防护机制。
- 模拟大规模并发和突发推理流量，以模拟生产规模的应用程序行为
- 通过自动化响应解析，测量每秒输入和输出 Token、TTFT、端到端延迟、吞吐量、并发性、带宽和基本响应准确性评分的实时指标。
- 全面的 REST API 支持自动化和 CI/CD 集成。



CyberFlood 高级推理配置

技术指标

LLM 模型	OpenAI/gpt-4.1，Ollama/llama
LLM 配置文件设置	TLS 加密选项
	身份验证
	提示词来源（用户输入或来自文件）
	对话模式
	Token 算法
提示词设置	文本、图像、关键词匹配、Token 估算
负载类型	模拟用户
关键指标	提示词 Token 总数
	发送的提示词总数
	响应词 Token 总数
	负载持续时间
	提示词 Token 持续时间
	执行的提示词
	每秒提示词 Token 数
	每秒提示词数
	每秒响应词 Token 数
	响应词 Token 使用量
	首次响应词 Token 时间（95/96/99 百分位）
	最后响应词 Token 时间（95/96/99 百分位）
支持的测试设备	CyberFlood Virtual
	CF400 和 CF30（已规划）

北京　　电话：+8610 8233 0055
上海　　电话：+8621 6859 5260
上海　　电话：+8621 2028 3588
（仅限 TeraVM 及 TM-500 产品查询）
深圳　　电话：+86 755 8869 6800
网站：　www.viavisolutions.cn

cyberflood-aiinterference-ds-hse-nse-zh-cn
30194862 900 1225