Создание агентной системы на основе LLM для тестирования веб-систем на проникновение

Автоматизация пентеста с помощью ИИ

Проблема

Ручное тестирование на проникновение

- Медленная скорость
- Нехватка экспертов
- Человеческий фактор

Актуальность

1. Рост сложности веб-приложений и количества уязвимостей
2. Традиционные средства автоматизации сильно ограничены
3. ИИ уже используется как атакующими, так и защитниками

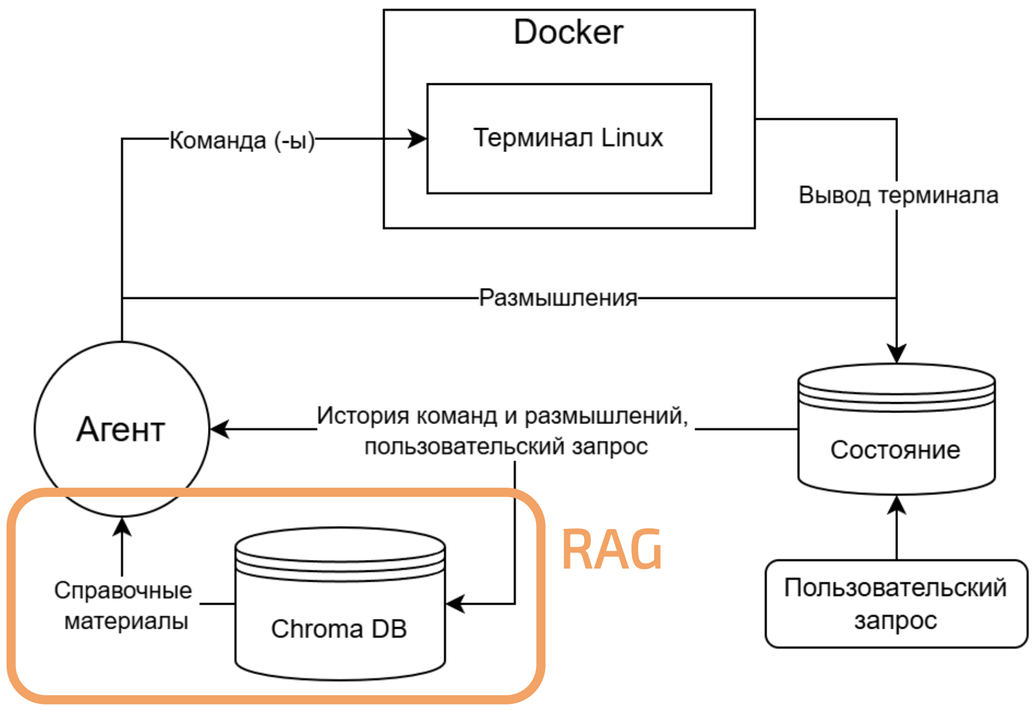
Цель

Создание агентной системы на основе большой языковой модели для автоматического тестирования веб-систем на проникновение.

Ход работы

Анализ предметной области → Проектирование архитектур → Реализация системы → Интеграция инструментов → Тестирование на датасетах → Анализ результатов

Архитектуры Single и Single-RAG



Single

- Один агент
- Линейный контекст
- Анализ, планирование и выполнение в одном

Single-RAG

- Один агент + знания
- Внешний контекст
- Больше когнитивная нагрузка модели

Архитектура P.H.A.N.T.O.M

- Несколько агентов
- Модель мира
- Стабильная работа на сложных задачах

Тестирование и сравнение

LLM	Mistral-7b-v0.3		Gemini-2.5-flash	
Архитектуры	Single	Single-RAG	Single	Single-RAG
Кол-во решенных задач	2 (4%)	0 (0%)	31 (62%)	36 (72%)

		Single	P.H.A.N.T.O.M
InterCode CTF	Кол-во реш. задач	45 (90%)	47 (94%)
	Ср. кол-во токенов	13578	19184
NYU CTF	Кол-во реш. задач	11 (19%)	18 (32%)
	Ср. кол-во токенов	109174	145542

Результаты

Разработана агентная система на основе LLM для автоматизированного решения CTF-задач. Эксперименты показали, что архитектура P.H.A.N.T.O.M работает стабильнее и эффективнее на сложных задачах, а RAG улучшает результаты крупных моделей, но может снижать их у малых.