Защита LLM от prompt-инъекций

Студент: Андриянов Эрик Вячеславович
Научный Руководитель: Сошников Д.В

CONTENTS

CONTENTS

Цель: Двусторонняя фильтрация запросов

Создать систему, которая защищает production-ботов российских компаний от утечек и дезинформации, обеспечивая:

< 20%
Доля успешных prompt-инъекций (база: 78%)

-70%
Снижение токсичности ответов (с ~0.72 до pan>

> 0.90
F1-метрика обнаружения PII

1.5-2 сек
Задержка полного цикла валидации

Prompt-инъекция — угроза №1 в 2025

По данным OWASP LLM Top 10, инъекции занимают первое место.

Быстрое исполнение

Атака выполняется за 42 секунды, regex-фильтры легко обходятся.

Высокие риски

Компании теряют данные, несмотря на рост эффективности LLM на 37%.

Массовые инциденты июля–августа 2025 подтверждают критичность перехода от keyword к semantic-защите.

Архитектура: Поток обработки запроса

Orchestration Layer выдаёт вердикт: ALLOW / BLOCK / MODIFY

Алгоритмы: Гибридный подход и скоринг

Rule-based

Regex, эвристики

ML

RF, XGBoost, DeBERTa

LLM-as-Judge

Embeddings

Semantic anomaly

→

Итоговый риск-скоринг

$$Risk = \alpha \cdot ML\_prob + \beta \cdot Rule\_score + \gamma \cdot Emb\_anomaly$$

Threshold-настройка для вердиктов:

BLOCK / REVIEW / ALLOW

Данные: 262k+ сэмплов

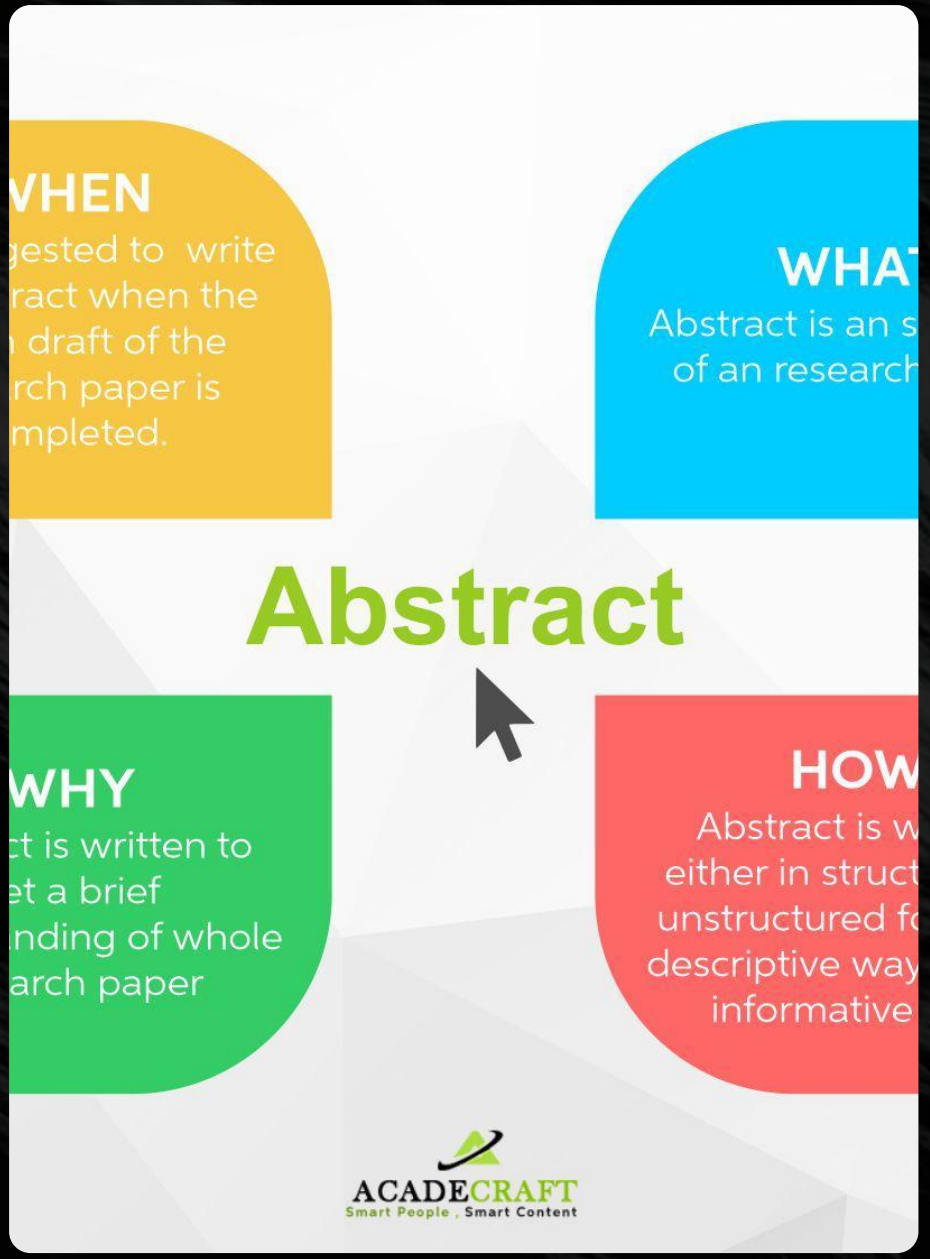
И синтетическая генерация для обучения.

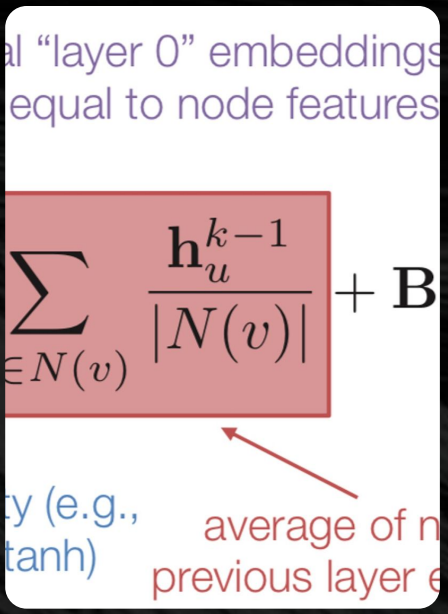
Публичные датасеты

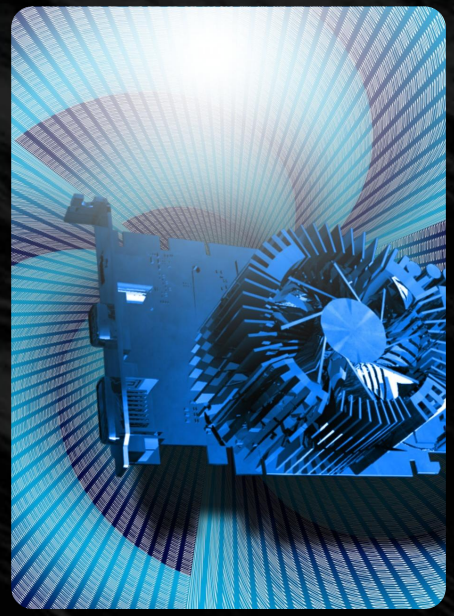
jayavibhav (262k), HackAPrompt (535k), CySecBench (12k)

Синтетика

Jailbreak DAN, role-play, paraphrase, Unicode-мутации







Существующие подходы:

Подход	Описание	Преимущества	Недостатки
Rule-based фильтрация	Regex, keyword filters, эвристические правила	Быстро (~50-100ms), легко интерпретируемо	ASR ~7-10%, легко обходятся (encoding, paraphrase)
ML-классификаторы (RF/XGBoost)	Random Forest, XGBoost на embeddings	Accuracy 84-90%, работа с дисбалансом	Требуют обучения, FPR ~1.7%
DeBERTa-детекторы	Fine-tuned DeBERTa-v3 для PI detection	Accuracy 97-99%, F1 0.94	Latency 200-500ms, требует GPU
Semantic Embeddings + Anomaly	LLM embeddings + autoencoder для anomaly	Обнаруживает novel attacks, контекстное	Сложность threshold-настройки
LLM-as-Judge	Вторая LLM для валидации запросов	Адаптируется к новым атакам	Latency 1-3 сек, высокая стоимость
Гибридный подход	Rule + ML + Embeddings + LLM-Judge	ASR <0.2%, многоуровневая защита	Сложность интеграции, latency 1.5-2 сек

Ключевые статистики и факты (2025)

Риск в OWASP

#1 позиция

Источник: OWASP LLM Top 10 2025

Рост атак

+540% за год

Источник: HackerOne Report 2025

Организации без защиты

97%

Источник: HackerOne Research

Инциденты июль-август 2025

4+ крупных утечек

Источник: Microsoft 365 Copilot, Cursor IDE, ChatGPT

Bug bounties AI

$2.1M (+339% YoY)

Источник: HackerOne 2025

ASR без защиты

7-10%

Источник: Baseline attacks

ASR с полной защитой

0.003%

Источник: Multi-layer defense

Ключевые выводы:

- Prompt injection - #1 риск по OWASP LLM Top 10 2025
- Рост атак на 540% за последний год
- 97% организаций не имеют адекватной защиты
- Крупные инциденты в июле-августе 2025 (Microsoft 365 Copilot, Cursor IDE, ChatGPT)
- $2.1M выплачено в bug bounties за AI-уязвимости

Гипотеза об эффективности решения

Основная гипотеза:

Гибридная архитектура с двусторонней фильтрацией (IVM + OVM) + multi-modal detection (Rule-based + ML + Embeddings + LLM-as-Judge) + risk scoring позволит достичь целевых метрик

Ожидаемые результаты:

Attack Success Rate	Токсичность	PII Detection F1
~~78% (baseline)~~ → <20%	~~0.72~~ → <0.18 (-70%)	~~~0.72~~ → >0.90

Latency (production)
~~N/A~~ → 1.5-2 сек

Обоснование: Каждый слой компенсирует слабости других: rule-based дает скорость, ML обнаруживает паттерны, embeddings ловят semantic anomalies, LLM-as-Judge адаптируется к novel attacks

Тестирование: Garak vs Llamator

Garak (NVIDIA)

- Плюсы: Множество готовых тестов, большое активное сообщество.
- Минусы: Слабый string-based детектор.

Llamator (ИТМО)

- Плюсы: Поддержка русского, многоступенчатые атаки.
- Минусы: Малое сообщество, базовый классификатор.

Цель тестирования: ASR < 20%, токсичность < 0.18, задержка < 2 сек.

Роадмап: От идеи к защите в проде

Октябрь 2025 → Май 2026

Октябрь-Ноябрь
Обзор литературы, сбор >100k данных

Декабрь-Январь
Rule-based компонент, baseline ML

Февраль-Март
Fine-tuning DeBERTa, тесты на бенчмарках

Апрель-Май
FastAPI wrapper, дашборд, документация

Май 2026: Защита ВКР