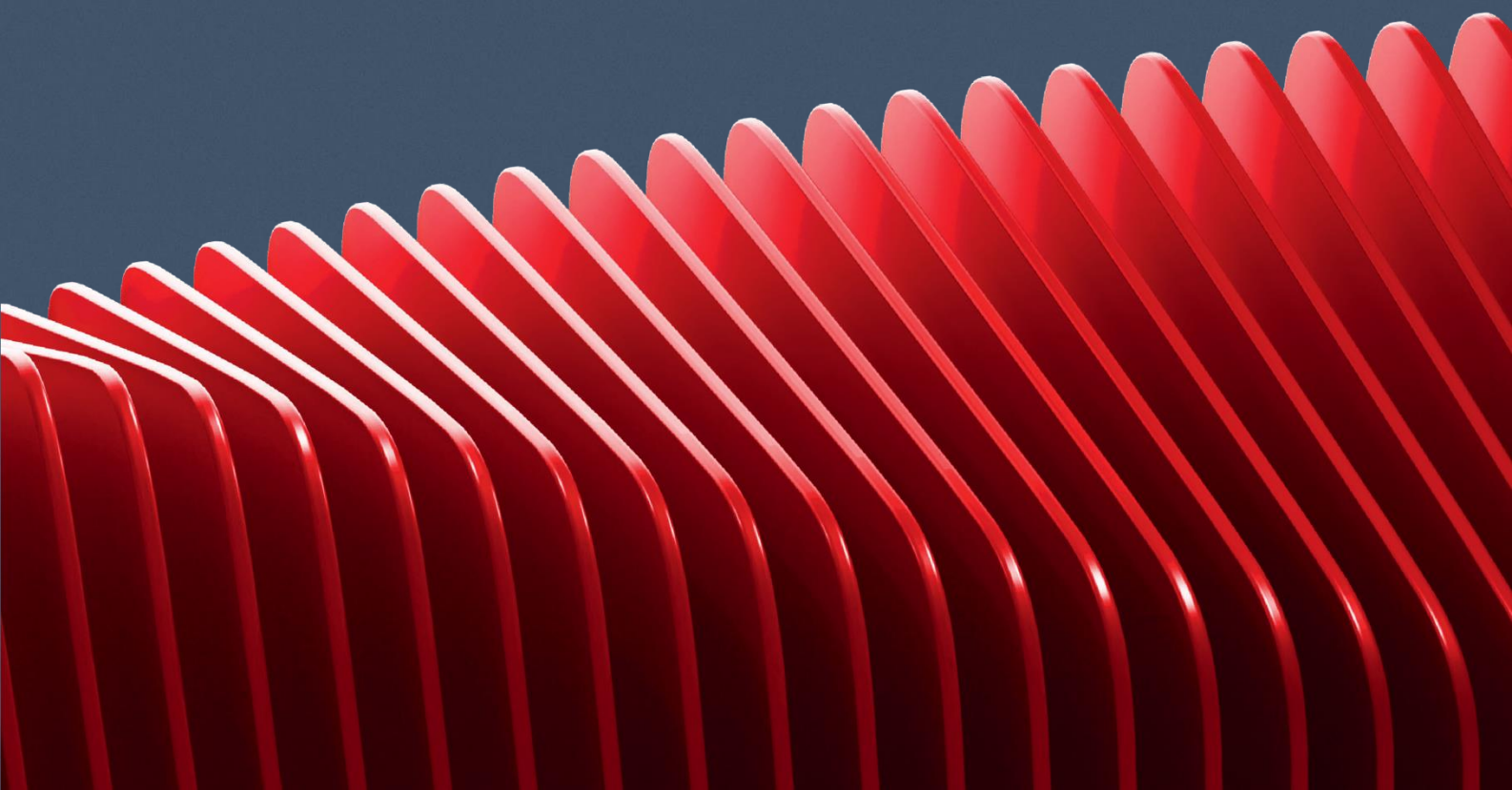
positive technologies

LLM Security

Чек-лист по безопасности больших языковых моделей



1 Реализуйте систему контроля доступа к API больших языковых моделей

Используйте API-ключи с разными правами в зависимости от того, какие доступы нужны пользователям, например, базируясь на существующих RBAC. Ограничивайте срок действия ключей и регулярно отслеживайте их – отзывайте неиспользуемые.

2 Внедрите MFA.

Особенно уделите внимание организации многофакторной аутентификации для доступов с расширенными правами (административные доступы, программные ключи).

3 Изолируйте LLM от интернета.

Размещайте сервисы в изолированных сетевых сегментах с учетом минимально необходимых доступов. Ограничьте подключения через API gateway.

4 Шифруйте данные, передаваемые в LLM.

Используйте надежные алгоритмы шифрования для трафика и для хранилищ. Ключи шифрования необходимо хранить в специализированных системах, которые приняты в компании.

5 Настройте маскирование или анонимизацию данных.

Маскирование или анонимизация чувствительных данных должны быть реализованы до обработки запросов. Выборочно аудируйте запросы к модели и ее ответы для внесения улучшений.

6 Организуйте логирование и хранение журнала логов.

Ведите детальные журналы и храните их в защищенных системах типа SIEM. Все логи должны содержать временные метки и идентификаторы пользователей. Ограничьте доступы к логам. Настройте правила, которые позволят мониторить подозрительные запросы – массовые запросы персональных данных, коммерческой тайны и другой чувствительной информации.

7 Настройте фильтрацию промпт-инъекций.

Создайте и интегрируйте правила, которые позволят обнаруживать зловредные запросы. Проводите регулярное тестирование установленных фильтров

8 Блокируйте ответы модели, содержащие чувствительную информацию.

Это позволит организовать защиту от утечки данных. Регулярно проводите аудит правил для блокировки и обновляйте их с учетом новых угроз.

9 Введите ограничения на длину, частоту запросов и на время ответа.

Ограничивайте запросы 1024 или 2048 токенов и настройте ratelimits на уровне API gateway, а также время ответа модели от 5 до 10 минут. Внедрите системы, позволяющие отслеживать аномальные скачки трафика.

10 Валидируйте источники данных.

Это позволит обеспечивать безопасность RAG. Проверяйте содержимое, добавляемое в RAG – их происхождение и содержиоме. Проводите автоматическое сканирование документов на вредоновные вложения (инъекции и JS-скрипты). Внедрите ограничения на используемые репозитории.

11 Настройте фильтрацию контекста, передаваемого в LLM (исключайте документы с маркированием).

Логируйте все использованные в генерации документы.

12 Изолируйте процессы обработки данных.

Введите запрет на сетевые вызовы, запусйкате процессы в изолированных средах (k8s) или serverless-средах. Мониторьте и реагируйте на аномальную активность (например, попытки доступа к паролям и учетных данным).

13 Используйте специализированные средства защиты.

LLM-фаерволы для блокировки prompt-injection и попыток джейлбрейка в запросах к модели. Обновляйте правила с учетом текущего ландшафта угроз, актуального ля вашей компании.

14 Разворачивайте LLM в контейнерах с минимальными правами.

Обеспечьте безопасность контейнерной инфраструктуры и применяйте security-профили.

15 Настройте дашборды для отслеживания метрик.

С помощью визуальных дашбордов реализуйте алертинг на множественные запросы и потребление избыточных ресурсов.

16 Обеспечьте юридическую и регуляторную безопасность.

Оцените, какие регуляторные меры действуют в вашей области бизнеса и обеспечьте удаление/маскирование/анонимизацию персональных и других чувствительных данных из датасета и запросов в соответствие им.

17 Проверяйте используемые модели.

Обеспечьте использования только доверенных репозиториев и мониторьте политики безопасности используемых LLM.

18 Проводите регулярное тестирование на проникновение.

Обеспечьте высокий уровень пентестов с учетом специфики окружения модели, интерфейсов и проверяйте на атаки, специфичные LLM (например, prompt-injection и др.) Тестируйте RAG на подверженность инъекциям.

19 Автоматизируйте деплой моделей и следите за обновлениями.

Настройте безопасность CI/CD и проверяйте целостность моделей перед развертыванием.

20 Создайте план реагирования на инциденты.

Учтите, что на каждый инцидент реагирование должно быть разным и проработайте разные сценарии. Проведите киберучения для тестировоания эффективности мер при работе во время инцидента.