КОНСОРЦИУМ
ИССЛЕДОВАНИЙ
БЕЗОПАСНОСТИ
ТЕХНОЛОГИЙ
ИСКУССТВЕННОГО
ИНТЕЛЛЕКТА

РАБОЧАЯ ГРУППА №1

ПО ПОДГОТОВКЕ ПРЕДЛОЖЕНИЙ В НОРМАТИВНЫЕ ПРАВОВЫЕ АКТЫ ДЛЯ РЕАЛИЗАЦИИ НАЦИОНАЛЬНОЙ СТРАТЕГИИ РАЗВИТИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ПЕРИОД ДО 2030 ГОДА

Алексей Сидорюк, отраслевой технический директор Positive Technologies

Известные проблемы и вызовы развития ИИ в России

Фокус деятельности рабочей группы

1 Специфические и динамические риски применения ИИ, слабо развитый рынок обеспечения доверия и защиты ML и LLM

Галлюцинации, отравление промптов, переобучение моделей и другие риски, которые слабо закрыты техническими средствами защиты и не всегда осознаются заказчиками

2 Массовое использование не проверенных Open Source моделей, открытых фреймворков, библиотек и датасетов

Для создания ИИ-решений в основном используются открытые IDE, библиотеки и датасеты, в которые могут быть встроены уязвимости и незадокументированные возможности

3 Отсутствие стандартов и механизмов проверки ИИ-решений для использования в КИИ и ГИС

В настоящий момент нет единых подходов по проверке ИИ-решений на уязвимости. Для заказчика решение зачастую представляет «черную коробку»

Использование ИИ в мошеннических целях

Дипфейки, подмена личности, социальный инжиниринг. ИИ сегодня активно используется злоумышленниками для обмана граждан

Дефицит кадров и компетенций, низкий уровень доверия населения

Дефицит кадров в ИИ составляет более 10 000 человек

Почти половина населения России по-прежнему не доверяет технологиям ИИ

Сложность контроля соблюдения этики. Важность правильной «картины мира» для GPT-решений

Нет объективных механизмов оценки соответствия ИИ-решений нормам этики

В GPT разные страны закладывают свои моральные ценности и «картину мира»

Цель и задачи регулирования

Цель

Сформировать набор предложений со стороны рабочей группы по изменению нормативно-правового ландшафта с целью обеспечения доверия и безопасности при использовании искусственного интеллекта для граждан, бизнеса и государства с целью реализации национальной стратегии развития ИИ 2030

Задачи

- Определить основные угрозы и риски применения искусственного интеллекта
- Сформировать требования к Реестру доверенных решений ИИ (в соответствии с предложениями РГ №3)
- Консолидировать предложения отрасли по стандартизации и регулированию ИИ для обеспечения доверия и безопасности технологий
- Разработать дорожную карту по изменениям НПА (внутри рабочей группы), определить исполнителей и согласовать подход с регуляторами
- Согласовывать материалы других рабочих групп

Ключевые принципы

1. «Не навреди»

Мы понимаем, что ИИ является стратегически важным комплексом технологий двойного назначения, нашей стране необходимо иметь технологический суверенитет в ней и высокие темпы развития. Цель нашей рабочей группы - за счет изменений регуляторной базы, стандартов и принципов этики обеспечить доверие и безопасность применения ИИ всех участников рынка - государства, бизнеса, граждан. Основной приоритет – формирование предложений для применения ИИ в государственном контуре и критической инфраструктуре

2. Комплексный подход к регулированию

Мы понимаем, что специфика технологий ИИ заключается в том, что они имитируют когнитивную деятельность человека, а все возможные сценарии их применения и ограничения невозможно зафиксировать в формате законодательных актов. Поэтому помимо предложений по изменению НПА могут быть сформированы изменения стандартов и принципов этики.

3. Открытость

Мы открыты к вступлению новых участников Консорциума в рабочую группу и к предложениям со стороны других рабочих групп

Уже есть инциденты информационной безопасности

1 Ray Framework vulnerability

Злоумышленники воспользовались CVE-2023-48022 – уязвимостью в фреймворке RAY, который используется для оркестрирования моделями машинного обучения. Воспользовавшись уязвимостью злоумышленники смогли исполнить код на хостах, где расположена модель.

https://www.oligo.security/blog/shadowray-attack-ai-workloads-actively-exploited-in-the-wild

2 VirusTotal Poisoning

Атака на систему анализа вредоносного ПО, где злоумышленник использует образец вредоносного ПО, пропускает его через метаморфический код и получает файлы, которые не запускаются, но распознаются антивирусом. Эти файлы попадают в датасет антивируса, отравляя его. В результате оригинальное вредоносное ПО становится труднее распознавать.
https://atlas.mitre.org/studies/AML.CS0002

3 Bypassing Cylance's AI Malware Detection. AI Cylance

Проведя анализ работы антивируса Cylance, изучив открытые источники, патенты и включив подробное ведение журнала злоумышленники выяснили, что ансамбль моделей (называемый "первой моделью") определяет вредоносное ПО, но его результаты могут отменяться второй моделью. Благодаря этому, злоумышленники смогли обойти основную модель, используя вторую модель.
https://atlas.mitre.org/studies/AML.CS0003

4 Compromised PyTorch Dependency Chain

Была скомпрометирована цепочка зависимостей. Вредоносная зависимость «torchtriton» на PyPI имеет то же имя, что и библиотека в репозитории PyTorch-nightly. Поскольку PyPI имеет приоритет в экосистеме Python, на машину попадает вредоносный пакет вместо PyTorch. Вредоносный «torchtriton» сканирует систему для получения базовой информации и крадет конфиденциальные данные.
https://atlas.mitre.org/studies/AML.CS0015

Какие бывают угрозы?

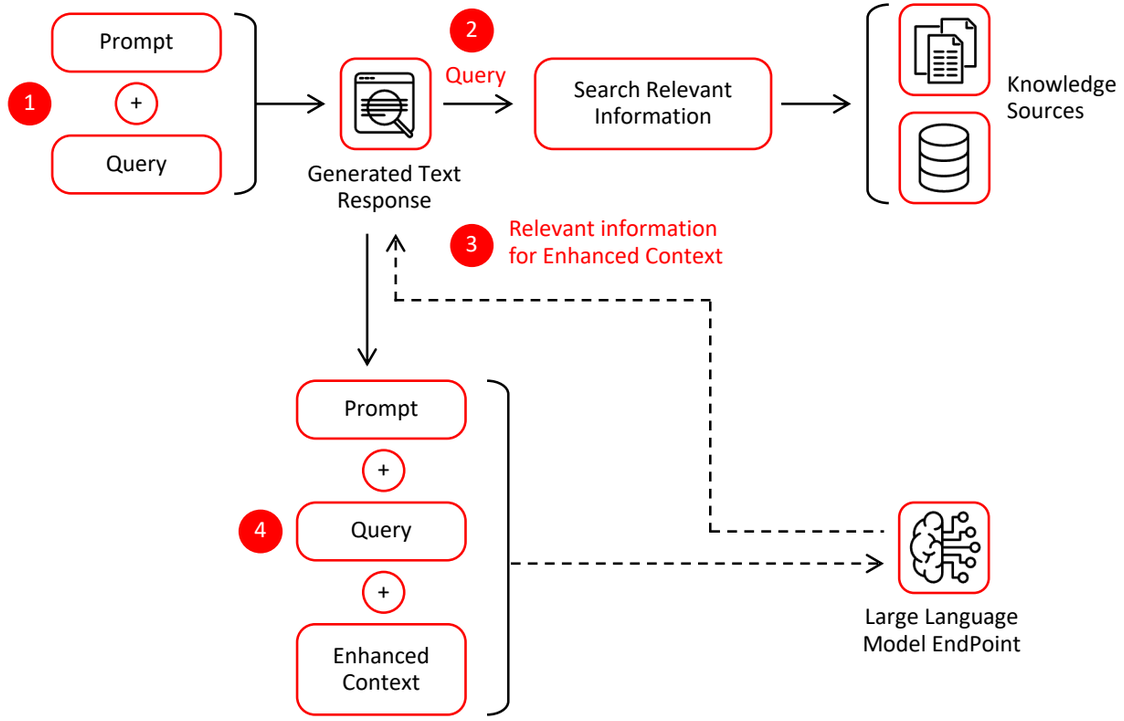
- Угрозы, связанные с моделью (ML и LLM)
- Угрозы, связанные с данными
- Угрозы, связанные с инфраструктурой
- Угрозы, связанные с процессами и с организационной частью

Угрозы по этапам жизненного цикла*

Описание бизнес-задачи и разработка ТЗ	Сбор данных	Предобработка и очистка	Разметка	Извлечение признаков	Обучение модели	Тестирование и экспорт	Разработка ПО	Мониторинг, эксплуатация и поддержка модели
Некорректное определение бизнес-задачи	Сбор низкокачественных, устаревших или нерепрезента-тивных данных	Недостаточная маскировка конфиденциальных данных	Манипулирование метками размеченных данных	Нелегитимное изменение извлеченных признаков	Несанкциониро-ванное изменение параметров модели	Нарушение требований тестирования модели	Использование компонентов с уязвимостями	Деградация модели или сервисов, представляемых на ее основе
Нарушение требований качества данных, надежности моделей и ПО	Нелегитимное изменение данных в процессе сбора	Несанкционированное изменение данных в процессе предобработки	Неполная или некачественная разметка данных	Использование вредоносной модели для извлечения признаков	Манипулирование алгоритмическими уязвимостями	Невозможно объяснить и интерпретировать результаты работы модели	Некорректная предобработка входных и постобработка выходных данных	Чрезмерное потребление ресурсов моделью
Отсутствие учета этических соображений	Сбор избыточно зависимых от одного источника данных	Некорректная реализация и уязвимости в предобработке данных	Предвзятая разметка	Неправильный выбор признаков	Переобучение модели	Небезопасный экспорт		Несанкциони-рованное использование результатов модели
	Сбор отравленных данных	Несанкциониро-ванный доступ к данным	Несанкциони-рованный доступ к данным	Несанкциониро-ванный доступ к признакам	Искажение распределения данных и деградация их качества в батчах при обучении	Внедрение бэкдоров в модель		Неограниченный доступ к ПО с моделью при отсутствии контроля
	Несанкциониро-ванный доступ к данным							

* Обезличенный пример перечня специфических угроз ИИ, возникающих на этапах жизненного цикла применения ИИ

Как это выглядит защита ИИ на практике (на примере GPT-моделей)



Анализ запросов

1. Обнаружение Jailbreaks и Prompt-инъекций
2. Обнаружение запросов на запрещенные темы
3. Обнаружение конфиденциальной информации в запросе

Анализ ответов

1. Обнаружение запрещенной информации
2. Обнаружение конфиденциальной информации в ответах модели (персональные данные, коммерческая тайна и т.д.)

Дальнейшие шаги

- [x] Утвержден состав рабочей группы
- [x] Разработана и согласована модель угроз искусственного интеллекта
- [x] Согласована концепция реестра отечественного ПО (подготовлена в рамках РГ №3)
- [ ] Модель угроз ИИ согласована со ФСТЭК
- [ ] Согласована методика и требования для тестирования решений на базе ИИ(подготовлена в рамках РГ №2)
- [ ] Сформирована дорожная карта по созданию НПА (внутри рабочей группы) и определены исполнители

Спасибо за внимание!