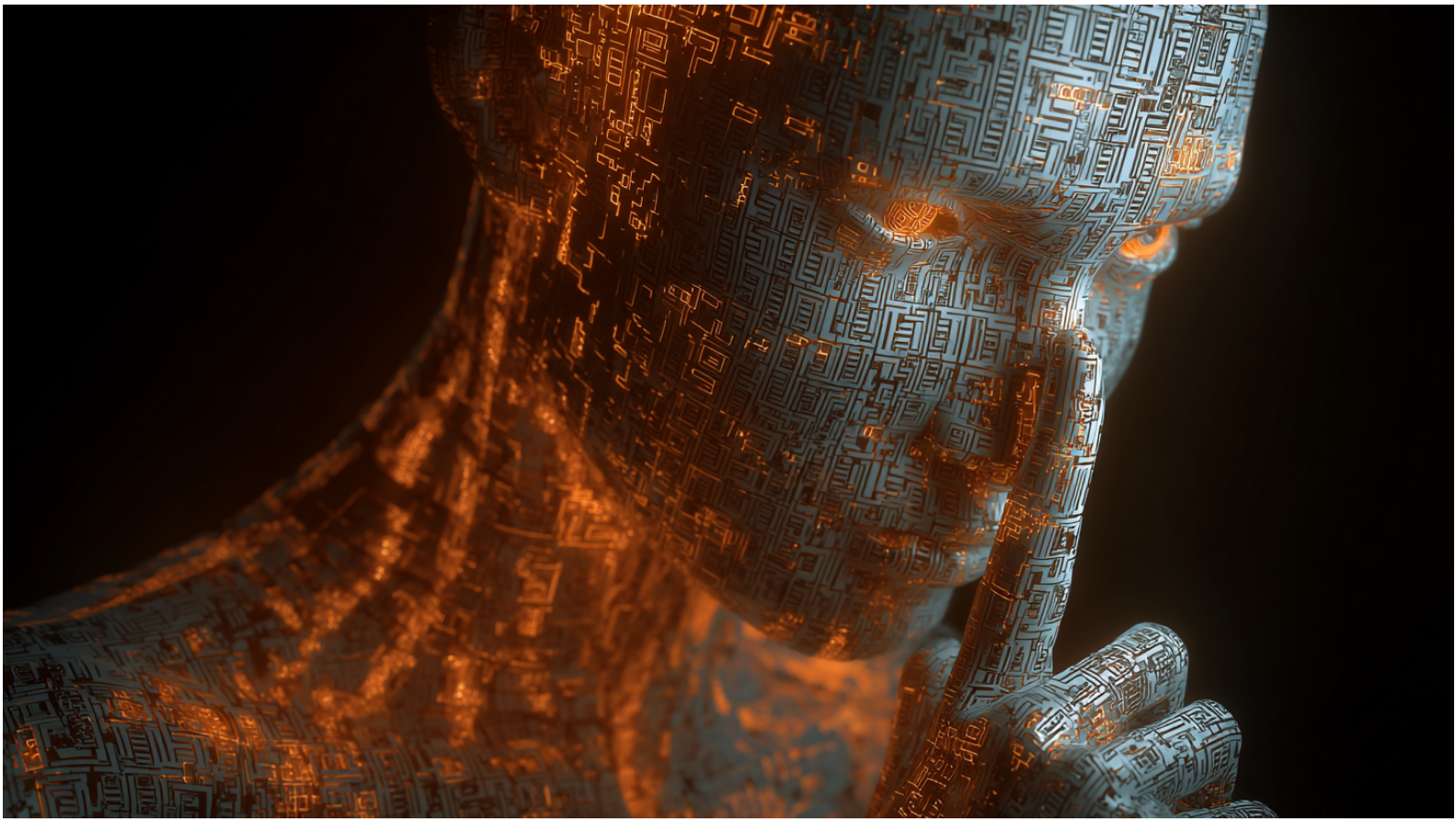
positive technologies

ПОЛИТИКА БЕЗОПАСНОСТИ LLM



Шаблон

Содержание

Перечень сокращений

Перечень сокращений, терминов и их определения приведены в таблице 1.

Таблица 1

Сокращение (термин)	**Расшифровка (определение)**
Атака	Преднамеренное действие или серия действий, направленных на нарушение конфиденциальности, целостности или доступности информационных систем, данных или ресурсов с целью получения несанкционированного доступа, повреждения или использования информации в ущерб владельцам систем
Безопасная разработка	Совокупность практик, инструментов, подходов и рекомендаций, направленных на повышение безопасности разрабатываемого и сопровождаемого ПО
Глубокое обучение	Раздел машинного обучения, основанный на использовании искусственных нейронных сетей с множеством слоёв (глубоких нейронных сетей), которые способны автоматически извлекать сложные признаки и представления из больших объёмов данных
Искусственный интеллект (ИИ)	Область исследований и технологий, направленная на создание систем, способных к автоматическому выполнению задач, требующих когнитивных функций человека, посредством моделирования и реализации алгоритмов машинного обучения, логического вывода и обработки естественного языка
Информационные технологии (ИТ)	Приемы, способы и методы применения средств вычислительной техники при выполнении функций сбора, хранения, обработки, передачи и использования данных
Машинное обучение	Область искусственного интеллекта, которая занимается разработкой алгоритмов и моделей,

Сокращение (термин)	Расшифровка (определение)
	позволяющих системам автоматически обучаться на данных, выявлять закономерности и принимать решения без явного программирования для каждой конкретной задачи
Модель ИИ	Математическая или алгоритмическая структура, созданная для имитации когнитивных функций человека, таких как обучение, распознавание образов, принятие решений и обработка информации, с целью выполнения определённых задач на основе входных данных
Большие языковые модели (Large Language Model, LLM)	Нейросетевой алгоритм, обученный на больших массивах данных с целью работы с естественного языка. Такие модели способны анализировать тексты и делать выводы на основе их содержания.
ОРД	Организационно-распорядительная документация
ПО	Программное обеспечение
Паттерн	(в контексте атаки) Повторяющаяся или типичная схема, модель или характерная последовательность действий злоумышленника, используемая для осуществления конкретного вида атаки на систему искусственного интеллекта.
Предвзятость	(в машинном обучении) систематическая ошибка или искажение в моделях ИИ или данных, приводящее к несправедливым, необъективным или некорректным результатам, обусловленным неравномерным представлением или неправильной настройкой обучающей выборки и алгоритмов
Репозиторий	Информационная система (ресурс), предназначенная, в том числе для размещения, хранения и распространения с использованием сети Интернет программного обеспечения для мобильных устройств

Сокращение (термин)	Расшифровка (определение)
Риск	Мера, учитывающая вероятность реализации угрозы и величину потерь (ущерба) от реализации этой угрозы
Тестирование на проникновение	Вид работ по выявлению (подтверждению) уязвимостей ПО или сервиса, основанный на моделировании (имитации) действий потенциального нарушителя
Уязвимость	Свойство информационных систем, которое можно использовать для нарушения функционирования ПО или несанкционированного доступа к его ресурсам
AI	(от англ. Artificial Intelligence) – искусственный интеллект

Общие положения

Настоящий документ описывает основные принципы и подходы, необходимые для обеспечения безопасности при разработке и использовании ПО с LLM

Цель данной политики — установить принципы и процедуры для безопасного и этичного функционирования больших языковых моделей (далее – LLM) в XXX организации. Основной целью настоящего документа является определение принципов, применяемых для безопасной разработки LLM, реализация которых позволяет обеспечить:

- минимизацию рисков и обеспечение этичного использования технологий;
- защиту данных, используемых для обучения моделей, включая анонимизацию и шифрование;
- проведение тестирования моделей на предмет их надежности, точности и безопасности;
- разработку мер по защите LLM от атак;
- регулярное обновление систем безопасности для защиты от новых угроз;
- применение методов, направленных на устранение предвзятости в данных и алгоритмах;
- установку систем мониторинга для отслеживания работы систем с использованием LLM после развертывания;
- возможность отзыва или корректировки ответов LLM в случае выявления негативных последствий.

Область применения

Данная политика применяется ко всем сотрудникам, подрядчикам и третьим лицам, участвующим в разработке и использовании ИИ-систем.

При построении процессов безопасной разработки ИИ-систем необходимо учитывать требования следующих нормативно-правовых актов и иных источников информации:

- Федеральный закон Российской Федерации (далее — РФ) от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации»;
- Федеральный закон РФ от 27 июля 2006 года №152-ФЗ «О персональных данных»;
- ГОСТ Р 71476-2024 «Искусственный интеллект. Концепции и терминология искусственного интеллекта»;
- ГОСТ Р 71539-2024 «Искусственный интеллект. Процессы жизненного цикла системы искусственного интеллекта»;
- ГОСТ Р 70889-2023 «Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных»;
- ГОСТ Р 56939-2024 «Защита информации. Разработка безопасного программного обеспечения. Общие требования» (далее — ГОСТ Р 56939-2024).

Основные положения безопасной разработки LLM

Принцип беспристрастности и недискриминации

Предвзятость в системах искусственного интеллекта может приводить к несправедливым, дискриминационным или предвзятым решениям и результатам, затрагивающим отдельные группы людей (например, по признакам пола, расы, возраста, социального положения, географического расположения и т.п.). Эта предвзятость возникает из данных, на которых обучаются модели: если данные неполны, нерепрезентативны, искажены или отражают существующие в обществе стереотипы и дискриминацию, модель их унаследует и воспроизведет. Беспристрастность означает, что решения, которые принимает LLM, должны приниматься исключительно на основе необходимости.

Чтобы противостоять этой проблеме, рекомендуется разрабатывать протоколы по использованию LLM, особенно в критически важных процессах. Также важно проверять результаты LLM на точность и корректность с данными в продуктовом контуре.

Для соблюдения принципа беспристрастности следует:

a. Проводить аудит и анализ обучающих и тестовых данных на предмет наличия смещений распределения и недостаточной репрезентативности.
b. Использовать разнообразные, качественные и репрезентативные наборы данных, отражающие все группы, на которые может повлиять система.
c. Регулярно тестировать разработанные модели на предмет наличия дискриминации или несправедливых результатов в отношении различных групп.

Прозрачность и объяснимость

Прозрачность означает, что пользователи должны понимать, как принимаются решения на каждом шаге работы модели. Необходимо чётко указывать, как система обрабатывает данные. Объяснимость в LLM означает, что система предлагает понятные человеку обоснования своих суждений и прогнозов.

Для соблюдения принципов прозрачности и объяснимости следует:

a. Документировать: архитектуру модели, используемые данные, методы обучения, ограничения и предположения.

Обеспечивать проверяемость ответов и объяснимость выводов. Модель должна ясно давать понять пользователю, где заканчивается её компетенция.

Защита данных и приватность

Защита данных и приватность при разработке LLM важны, поскольку системы искусственного интеллекта работают с большими объемами данных, которые могут содержать конфиденциальную, персональную или чувствительную информацию. Ненадлежащая защита этих данных создает риски, включающие: утечки информации, нарушение приватности пользователей, незаконное использование данных, потерю доверия, а также юридические последствия в случае несоответствия законодательным требованиям.

Для соблюдения принципов прозрачности и объяснимости следует:

a. Строго соблюдать законодательство о защите данных и персональной информации на всех этапах жизненного цикла данных, используемых в ИИ (сбор, хранение, обработка, обучение).
b. Шифровать данные как при хранении, так и при передаче.
c. Применять методы анонимизации, псевдонимизации или агрегирования данных, чтоб устранить или уменьшить прямую связь данных с конкретными лицами.
d. Внедрять строгие меры контроля доступа к данным и моделям.
e. Разработать и соблюдать политики хранения и безопасного удаления данных после того, как они перестают быть необходимыми.

Тестирование, валидация, верификация

Сложность LLM и их зависимость от данных могут приводить к неожиданному поведению, ошибкам или снижению производительности на наборе данных, отличном от обучающего. Надлежащее тестирование и валидация необходимы для снижения этих рисков и обеспечения того, что модель выдает корректные результаты, является надежной и безопасной в реальных условиях.

При комплексном тестировании LLM следует:

a. Проверять легитимность данных на всех этапах разработки модели. Нужно тщательно проверять поставщиков данных и сравнивать выводы модели с доверенными источниками.
b. Использовать методы обнаружения аномалий для фильтрации вредоносных данных.
c. Применять управление версиями данных, чтобы отслеживать изменения в наборах данных и выявлять манипуляции.
d. Тестировать устойчивость модели с помощью AI Red Teaming и техник противодействия.
e. Отслеживать потери на этапе обучения и анализировать поведение модели на наличие признаков отравления.
f. При использовании Retrieval-Augmented Generation (RAG) организовать разграничение доступа.
g. Валидировать поведение на независимых и разнообразных наборах данных, которые максимально репрезентативны по отношению к реальным данным, с которыми система столкнется во время эксплуатации. Нужно тщательно проверять поставщиков данных и сравнивать выводы модели с доверенными источниками.
h. Следует использовать системы обнаружения аномалий, которые отслеживают взаимодействие LLM в реальном времени для выявления деградации, дрейфа данных или аномального поведения, требующего переобучения или корректировки модели.

Устойчивость к атакам

LLM, как и любое сложное программное обеспечение, подвержены различным типам атак, которые могут быть направлены на нарушение их корректной работы, манипулирование результатами, кражу конфиденциальных данных а также нанесение вреда. Однако LLM имеют специфические уязвимости, связанные с их зависимостью от данных и сложностью внутренних механизмов принятия решений.

Для обеспечения устойчивости к атакам следует применять следующие меры:

a. Выявить потенциальные вектора атак, специфичные для LLM и области ее применения. Это может включать атаки на обучающие данные, на инфраструктуру, на конвейеры данных и на результирующие решения.

b. Вводить ограничения поведения модели. Нужно заранее сформулировать жёсткие правила относительно ролей, возможностей и ограничений модели. Модель следует настраивать так, чтобы она игнорировала команды, противоречащие базовым инструкциям.
c. Следует определять критерии формата ответа.
d. Производить фильтрацию входного и выходного контента. Следует определить категории опасной или конфиденциальной информации и разработать правила для её выявления. Нужно внедрять семантические фильтры и регулярные выражения, чтобы вовремя выявлять потенциально вредоносные запросы или ответы.
e. Следовать принципу минимальных привилегий. Модели следует выдавать только те права, которые действительно необходимы для выполнения поставленной задачи. Расширенную функциональность нужно обрабатывать на уровне приложения, а не самой модели.
f. Обеспечивать человеческий контроль при выполнении действий, которые могут иметь серьёзные последствия. Это позволит остановить опасные процессы до их завершения.
g. Реализовать процесс отделения и маркировки внешнего контента. Нужно явно помечать непроверенные данные и не позволять им напрямую влиять на внутреннюю логику запросов или ответы модели.
h. Обучение персонала. Регулярные тренинги для разработчиков и пользователей повысят осведомлённость о потенциальных угрозах и методах защиты.
i. Внедрить системы мониторинга для выявления подозрительной активности или изменения в поведении модели, которые могут указывать на атаку.

Мониторинг и отзыв

Мониторинг важен для обнаружения возникающих угроз и изменений в поведении. Кроме того, только в реальной эксплуатации могут проявиться некоторые этические проблемы, проблемы безопасности или негативные социальные последствия, которые не были выявлены на этапе тестирования. Принцип мониторинга и отзыва обеспечивает постоянный надзор за работой системы и возможность своевременного реагирования на возникающие проблемы.

Для соблюдения принципа мониторинга и отзыва следует:

a. Необходимо постоянно наблюдать за аномалиями или отклонениями от ожидаемого поведения в реальном времени или с высокой периодичностью.
b. Нужно следить, чтобы ответы модели не содержали конфиденциальный или манипулированный контент.
c. Настроить алерты, срабатывающие при обнаружение личной информации, финансовых данных или других запрещённых типов вывода или при выявлении аномалий. Если ответ помечается как запрещённый, его к нему необходимо предпринять возможные действия: заблокировать и заменить запасным вариантом, отправить на доработку, передать на рассмотрение человеку.
d. Создать процессы и команды для оперативного расследования причин снижения производительности, выявления предвзятости, инцидентов безопасности или других проблем, обнаруженных через мониторинг или обратную связь.
e. Обеспечить возможность возвращения к предыдущей версии модели или системы, если выявлены критические проблемы.
f. Использовать информацию, полученную в результате мониторинга и анализа обратной связи, для улучшения будущих версий модели, совершенствования обучающих данных, доработки процессов тестирования и валидации, а также повышения общей надежности и безопасности системы.

Безопасность инфраструктуры

Безопасность инфраструктуры – это принцип обеспечения защиты инфраструктурных компонентов, используемых для разработки и эксплуатации систем с искусственным интеллектом, с целью минимизации рисков и предотвращения возможных угроз.

Для реализации этого принципа необходимо:

a. Провести подготовительные работы по обеспечению безопасности инфраструктуры, учитывая принципы безопасности и меры по снижению рисков.
b. Настроить конфигурацию безопасности инфраструктуры в соответствии с требованиями, установленными в рамках проектирования архитектуры.

c. Регулярно (не реже чем два раза в год) проверять соответствие конфигурации требованиям безопасности, а при выявлении отклонений — задокументировать и реализовать корректирующие меры.
d. Определить ролевую модель доступа, включающую роли, права и ресурсы, обеспечивающие контроль доступа к инфраструктуре.
e. Обеспечить разделение инфраструктурных сред для различных стадий жизненного цикла разработки (разработка, тестирование, сборка, развертывание), а также разделение пользователей на изолированные сетевые группы в соответствии с ролевой моделью.
f. Настроить процесс разработки так, чтобы фиксировать процессы сборки и генерации подписи для подтверждения происхождения результатов.
g. Определить элементы конфигурации, связанные с разрабатываемой системой, подлежащие резервному копированию, и организовать централизованное автоматическое резервное копирование репозиториев исходного кода и зависимостей.
h. Внедрить регулярный анализ безопасности инфраструктуры, включая оценку уязвимостей, проведение тестов на проникновение и аудит конфигураций.

Повышение экспертизы

Повышение экспертизы в области разработки LLM и обеспечении информационной безопасности в процессе разработки— это принцип, направленный на систематическое развитие знаний и навыков сотрудников, обеспечивающее высокий уровень компетентности в области безопасной разработки.

Для реализации этого принципа необходимо:

a. Регулярно проводить обучение сотрудников, включающее теоретические знания и практические навыки, чтобы поддерживать и повышать уровень экспертизы.
b. Организовывать периодические (не реже одного раза в год) обучающие мероприятия, охватывающие новые практики безопасной разработки ПО для разработки систем искусственного интеллекта, актуальные виды атак и уязвимостей, современные методы и инструменты безопасности, а также нормативные требования. В рамках обучения должны проводиться практические занятия по применению полученных знаний.

c. Разработать и поддерживать в актуальном состоянии внутренний портал по безопасности, содержащий информацию о практиках безопасной разработки, нормативных требованиях, типах угроз и внутренней документации (внутренние ОРД, описание разрабатываемых систем и т.д.).
d. Внедрить механизмы стимулирования работников к повышению квалификации, такие как предоставление возможности посещения специализированных курсов и конференций, а также выплаты премий за успешное прохождение сертификаций по безопасности разработки.

Ответственность за соблюдение политики и ее пересмотр

Все сотрудники несут ответственность за соблюдение данной политики. Нарушения политики могут привести к дисциплинарным мерам.

Данная политика подлежит регулярному пересмотру и обновлению в соответствии с изменениями в технологиях, законодательстве и этических нормах.

ПОЛЕЗНЫЕ РЕСУРСЫ POSITIVE TECHNOLOGIES

- Сайт Positive Technologies - https://ptsecurity.com/ru-ru/
- Telegram-канал Positive Technologies - https://t.me/Positive_Technologies
- Сайт Positive Hack Days - https://phdays.com/ru/
- Сайт Standoff 365 - https://standoff365.com
- Издание Positive Research - https://ptresearch.media
- Telegram-канал ESCalator - https://t.me/ptescalator

Вас взломали? Мы поможем!

Перейдите по QR-коду

https://www.ptsecurity.com/

+7 495 744 01 44

pt@ptsecurity.com

t.me/Positive_Technologies