ИНТЕЛЛЕКТУАЛЬНЫЙ СЕТЕВОЙ ЭКРАН ДЛЯ ЗАЩИТЫ ИИ-ПРИЛОЖЕНИЙ SOLIDWALL AI SECURITY GATEWAY

Описание функциональных характеристик

Листов: 15

2026 г.

АННОТАЦИЯ

Данный документ содержит описание функциональных характеристик программного обеспечения и информацию, необходимую для установки программного обеспечения.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

1.1 Область применения

Область применения интеллектуального сетевого экрана для защиты ИИ-приложений SolidWall AI Security Gateway (далее – SolidWall AI Security Gateway) -– мониторинг состояния защищённости ИИ-приложений.

1.2 Краткое описание возможностей

SolidWall AI Security Gateway является интеллектуальным межсетевым экраном прикладного уровня и предназначен для мониторинга трафика ИИ-приложений на наличие интернет-угроз.

SolidWall AI Security Gateway выполняет следующие функции:

- анализ трафика веб-приложений с искусственным интеллектом и обнаружение атак (вторжений);
- блокирование попыток сетевых атак при работе с веб-приложениями с искусственным интеллектом;
- контроль доступа к функциям и журналирование действий пользователей защищаемых приложений с искусственным интеллектом;
- обнаружение подозрительной активности пользователей веб-приложений с искусственным интеллектом;
- автоматическая настройка защитных механизмов «SolidWall AI Security Gateway», версия 1.0 под конкретное защищаемое приложение с искусственным интеллектом (обучение);
- контроль и фильтрацию трафика, включая запросы пользователя к языковым моделям;
- идентификацию и аутентификацию пользователей, являющихся работниками оператора;
- регистрацию событий безопасности (аудит);
- бесперебойное функционирование и восстановление;
- тестирование и контроль целостности;
- управление (администрирование);
- взаимодействие с другими средствами защиты информации.

2 НАЗНАЧЕНИЕ И УСЛОВИЯ ПРИМЕНЕНИЯ

SolidWall AI Security Gateway является интеллектуальным межсетевым экраном прикладного уровня и предназначен для мониторинга трафика ИИ-приложений на наличие интернет-угроз.

2.1 Функциональные возможности SolidWall AI Security Gateway

SolidWall AI Security Gateway выполняет следующие функции:

- анализ трафика веб-приложений с искусственным интеллектом и обнаружение атак (вторжений);
- блокирование попыток сетевых атак при работе с веб-приложениями с искусственным интеллектом;
- контроль доступа к функциям и журналирование действий пользователей защищаемых приложений с искусственным интеллектом;
- обнаружение подозрительной активности пользователей веб-приложений с искусственным интеллектом;
- автоматическая настройка защитных механизмов «SolidWall AI Security Gateway», версия 1.0 под конкретное защищаемое приложение с искусственным интеллектом (обучение);
- контроль и фильтрацию трафика, включая запросы пользователя к языковым моделям;
- идентификацию и аутентификацию пользователей, являющихся работниками оператора;
- регистрацию событий безопасности (аудит);
- бесперебойное функционирование и восстановление;
- тестирование и контроль целостности;
- управление (администрирование);
- взаимодействие с другими средствами защиты информации.

Основные функции системы:

- защита от основных видов атак на AI-приложения из перечней OWASP LLM Top-10 Risks и OWASP Top-10 for Agentic Applications, в том числе от атак инъекций запросов к LLM (prompt injection), нарушения ограничений LLM (jailbreak), утечек чувствительной информации;
- защита от основных видов атак на веб-приложения из перечня OWASP Top-10, от логических атак на AI-приложения, в том числе от атак на механизмы аутентификации и контроля сессий, а также атак на бизнес-логику;
- полнофункциональная защита API (функционал API Gateway) с возможностью создания и корректировки моделей, вручную или автоматически, включая определение форматов данных и методов взаимодействия, не поддерживаемых стандартом Open API;

– отражение переборных атак и ботов, сочетающее позитивную модель, рейт-лимитинг, анализ поведения пользователей при обращении к AI-специфичной функциональности приложения;

– противодействие «умному DoS-у» (целевым атакам на уровне приложений (L7) на исчерпание ресурсов LLM моделей), учитывающее особенности функционирования больших языковых моделей;

– контроль действий пользователей на основе анализа и хранения легитимных транзакций, а также использования механизмов контроля бизнес-логики (анализ последовательностей обращений к AI-специфичным функциям, UBA и т. п.).

Функциональные возможности SolidWall AI Security Gateway определены следующим составом:

– возможность осуществлять мониторинг сетевого трафика для отправителей информации, получателей информации и всех операций передачи контролируемой SolidWall AI Security Gateway информации к веб-серверу и от веб-сервера;

– возможность обеспечить, чтобы в сетевом экране мониторинг распространялся на все операции перемещения через SolidWall AI Security Gateway информации к веб-серверу и от веб-сервера;

– возможность поддержки контроля и анализа запросов и ответов по протоколу передачи гипертекста определенных версий;

– возможность поддержки контроля и анализа сообщений, отправляемых веб-браузером веб-серверу и содержащих текстовый контент определенных кодировок и нетекстовый контент определенных типов (изображения, аудиоинформация, видеоинформация, программы);

– возможность поддержки контроля и анализа специальных маркеров взаимодействия (куки) определенных типов, отправляемых веб-сервером веб-браузеру и возвращаемых веб-браузером веб-серверу, содержащих персонифицированную информацию сеанса взаимодействия пользователя с веб-сервером, основываясь на определенных атрибутах куки;

– возможность помечать разрешённым информационный поток, базируясь на устанавливаемом администратором SolidWall AI Security Gateway наборе правил фильтрации, основанных на идентифицированных атрибутах;

– возможность помечать запрещённым информационный поток, базируясь на устанавливаемом администратором SolidWall AI Security Gateway наборе правил фильтрации, основанных на идентифицированных атрибутах;

– возможность осуществлять проверку наличия фрагментов мобильного кода в запросах пользователей к сайту и (или) иному ИИ-приложению на ввод данных путем поиска в таких запросах определенных фрагментов регулярных выражений (тегов, команд в формате языков мобильного кода), используемых при инициализации мобильного кода или выполнения нежелательных действий;

– возможность помечать запрещённым информационный поток, если в нём обнаружен запрос пользователя к сайту и (или) иному ИИ-приложению на ввод данных, содержащий выявленный мобильный код;

– возможность помечать разрешённым информационный поток, основываясь на результатах проверок;

– возможность помечать запрещённым информационный поток, основываясь на результатах проверок;

– возможность помечать разрешённым информационный поток, если значения атрибутов безопасности, установленные взаимодействующими средствами защиты информации для контролируемого сетевого трафика, указывают на отсутствие нарушений безопасности информации;

– возможность помечать запрещённым информационный поток, если значения атрибутов безопасности, установленные взаимодействующими средствами защиты информации для контролируемого сетевого трафика, указывают на наличие нарушений безопасности информации;

– возможность поддержки контроля и анализа фрагментированных запросов и ответов при доступе к веб-серверу;

– возможность поддержки контроля и анализа запросов и ответов при доступе к веб-серверу, подвергшихся сжатию;

– возможность поддержки отправления и получения данных пользователя информационной системы (в том числе куки – фрагменты информации, содержащие данные пользователя) способом, защищенным от несанкционированного раскрытия;

– возможность поддержки отправления и получения данных пользователя информационной системы (в том числе куки – фрагменты 9 информации, содержащие данные пользователя) способом, защищенным от нарушения целостности;

– возможность помечать запрещённым неразрешенного информационного потока по протоколу передачи гипертекста одним или несколькими способами: блокирование запроса по протоколу передачи гипертекста;

– пометка запрещённым взаимодействия с конкретным сетевым адресом;

– пометка запрещённым сессии на уровне конкретного приложения;

– пометка запрещённым взаимодействия на уровне конкретного пользователя приложения;

– отправка управляющего сигнала на иной SolidWall AI Security Gateway для пометки запрещённым неразрешенного информационного потока;

– возможность уведомления (оповещения) администратора SolidWall AI Security Gateway о выполненной пометке запрещённым неразрешенного информационного потока по протоколу передачи гипертекста;

– возможность снятия пометки о блокировке информационных потоков администратором SolidWall AI Security Gateway, установленной ранее;

– возможность поддержки виртуализации внешнего представления приложений веб-сервера на уровне трансляции сетевых портов;

– возможность уведомления (оповещения) пользователя информационной системы о выполненной пометке о блокировке неразрешенного информационного потока по протоколу передачи гипертекста;

– возможность поддержки виртуализации внешнего представления приложений веб-сервера на уровне трансляции унифицированных идентификаторов ресурсов;

– возможность определения виртуализированных (видимых для веб-браузера) сетевых портов приложений и их сопоставления с реальными (видимыми для веб-сервера) сетевыми портами приложений;

– возможность определения виртуализированных (видимых для веб-браузера) идентификаторов ресурсов и их сопоставления с реальными (видимыми для веб-сервера) унифицированными идентификаторами ресурсов;

– возможность обеспечения перехода в режим аварийной поддержки, который предоставляет возможность возврата SolidWall AI Security Gateway к штатному режиму функционирования;

– возможность тестирования (самотестирования) функций безопасности SolidWall AI Security Gateway (контроль целостности исполняемого кода SolidWall AI Security Gateway);

– возможность согласованно интерпретировать управляющие команды, атрибуты сетевого трафика и иные данные, получаемые от взаимодействующих с SolidWall AI Security Gateway средств защиты информации других видов;

– поддержка правил интерпретации данных, получаемых от взаимодействующих с SolidWall AI Security Gateway средств защиты информации других видов;

– возможность завершения работы или восстановления (для предусмотренных сценариев сбоев) штатного режима функционирования SolidWall AI Security Gateway;

– возможность читать информацию из записей аудита уполномоченным администраторам;

– возможность выбора совокупности событий, подвергающихся аудиту, из совокупности событий, в отношении которых возможно осуществление аудита;

– возможность оповещения уполномоченных лиц о критичных видах событий безопасности, в том числе сигнализация о событиях безопасности, связанных с обнаружением неразрешенных информационных потоков по протоколу передачи гипертекста;

– возможность выборочного просмотра данных аудита (поиск, сортировка, упорядочение данных аудита);

– возможность аутентификации администратора SolidWall AI Security Gateway до разрешения любого действия (по администрированию), выполняемого при посредничестве SolidWall AI Security Gateway от имени этого администратора;

– возможность идентификации администратора SolidWall AI Security Gateway до разрешения любого действия (по администрированию), выполняемого при посредничестве SolidWall AI Security Gateway от имени этого администратора;

– возможность осуществления идентификации пользователя до разрешения передачи через SolidWall AI Security Gateway информационного потока, ассоциированного с этим пользователем, к веб-серверу (от веб-сервера);

– возможность осуществления аутентификации пользователя до разрешения передачи через SolidWall AI Security Gateway информационного потока, ассоциированного с этим пользователем, к веб-серверу (от веб-сервера);

– поддержка определенных ролей по управлению SolidWall AI Security Gateway;

– анализ трафика ИИ-приложений и обнаружение атак (вторжений);

– пометка попыток сетевых атак при работе с ИИ-приложениями;

– информирование о проведении следующего вида атак на ИИ-приложение:

- всевозможные виды инъекций (SQL injection, OS injection, RCE, XPath-injection, XXE);
- реализация атак на обратный путь в директориях и на включение удаленных или локальных файлов (Directory traversal, Remote/Local File Inclusion);
- межсайтовый скриптинг (XSS);
- межсайтовая подделка запросов (CSRF);
- реализация атак, эксплуатирующих небезопасные настройки ИИ-приложений (security misconfiguration);
- реализация переборных атак (bruteforce);
- «умный» DoS (application-level DoS);
- реализация атак, эксплуатирующих недостатки системы аутентификации (фиксация сессии, кража сессии, отсутствие таймаута и т. п.);
- реализация атак на механизмы авторизации (Insecure Direct Object References, Missing Function Level Access Control);
- автоматизированный обход приложения, массовое скачивание информации, использование сканеров уязвимостей (web scraping, automation);

– обнаружение подозрительной активности пользователей ИИ-приложений.

2.2 Преимущества SolidWall AI Security Gateway

SolidWall AI Security Gateway имеет следующие преимущества:

– Интерфейс управления:

- Управление осуществляется по протоколу HTTP с использованием специализированного веб-интерфейса.
- В целях интеграции с другими решениями, управление может осуществляться также с использованием REST API.
- Обеспечение возможности мониторинга и управления конфигурациями для нескольких защищаемых приложений из общего интерфейса.

- Редактирование конфигурации подсистемы захвата трафика (обратный-прокси) с использованием графического интерфейса.
- Возможность загрузки SSL-сертификатов в графическом интерфейсе Системы.
- Поддержка регулярных выражений (wildcard) для доменных имен защищаемых приложений в настройках защиты.
- Ограничение доступа пользователей только к определенным группам приложений (режим Multitenancy).
- Поддержка графического отображения и редактирование правил защиты AI-приложений, раздельно для каждого типа атак на LLM из перечней OWASP LLM Top-10 Risks и OWASP Top-10 for Agentic Applications, с возможностью настройки реакции на каждый тип атаки независимо – блокировка, маркирование потенциально вредоносных запросов, отправка на капчу и др.
- Веб-интерфейс обеспечивает версионность каждой из конфигурационных настроек системы: в т. ч. базовых конфигурационных настроек, настроек модулей анализа, моделей нормального функционирования приложений, правил принятия решений.
- Интерфейс поддерживает механизмы аутентификации и авторизации пользователей. Ролевая модель предусматривает разграничение доступа для трех групп: администраторы (можно выполнять любые действия), аналитики (можно менять только настройки безопасности), операторы (только мониторинг и работа с событиями безопасности).
- Действия пользователей веб-интерфейса, включая внесенные изменения, сохраняются во внутреннем журнале аудита системы.

– Работа с использованием негативных методов обнаружения атак на AI-приложения:

- Наличие предобученного классификатора вредоносных запросов Yandex LLM Censor.
- Наличие набора встроенных сигнатур для атак из перечня OWASP LLM Top-10 Risks в комплекте поставки.
- Создание аномальных последовательностей (цепочек) действий при взаимодействии с AI-эндпоинтами, для моделирования нежелательного поведения.

– Противодействие переборным атакам и повторяющимся запросам:

- Возможность ограничения количества запросов, направляемых приложению в единицу времени.
- Возможность индивидуальной настройки параметров противодействия переборным атакам для отдельных эндпоинтов, реализующих функциональность AI агентов или использующих LLM.
- Детектирование бот-активности по аномальному превышению медианного времени задержки ответа AI-приложения по сравнению с историческими данными.

– Работа на основе позитивных моделей приложения

- Наличие готовых моделей валидации запросов и ответов на наличие чувствительной информации для типового AI-приложения в комплекте поставки.
- Возможность гибкой настройки различных типов моделей для каждого из защищаемых AI-приложений, в том числе:
  - модели валидации протокола HTTP;
  - модели синтаксического анализа запросов и ответов с поддержкой различных видов сжатия, кодирования и способов передачи данных с различными уровнями вложенности (в частности, XML, JSON, BASE64, GZIP, SOAP, Zstd, Brotli, Protocol Buffers);
  - модели проверки соответствия XML-подобных форматов данных заданным схемам XSD, загружаемым в виде файлов в интерфейсе AI Security Gateway;
  - модели источников – определение характеристик источника на основе параметров запроса;
  - модели определения логических действий (бизнес-действий) в приложении, параметров логических действий и их значений, последовательностей действий (цепочек), проверки успешности действий;
  - модели защиты от переборных атак и атак типа «умный DoS» на уровне отдельных логических действий и произвольных параметров действия;
  - модели пользователей, их идентификации, аутентификации и контроля сессий.
- Наличие готовых моделей валидации протокола HTTP и синтаксического анализа запросов для типового веб-приложения в комплекте поставки.
- Наличие автоматического алгоритма оценки отклонения параметров действий AI-приложения от статистической нормы.

– Обнаружение аномалий и значимых событий:

- Возможность обнаружения аномалий или чувствительных данных как в HTTP-запросах, так и в HTTP-ответах.
- Возможность обнаружения аномалий AI моделей следующих типов: инъекции вредоносных запросов к LLM (prompt injection), нарушение ограничений LLM (jailbreak), фишинг и социальная инженерия (deception), токсичный и нежелательный контент (toxic content), признаки мошенничества (fraud&scam), утечка чувствительных данных (раздельно персональные данные, корпоративная политика, финансовые данные, потенциальное нарушение закона, медицинские данные), и другие.
- Возможность обнаружения аномалий работы приложения на основе сопоставления значений параметров HTTP-запросов/ответов с сигнатурами атак на AI-приложения и веб-приложения.

- Возможность обнаружения аномалий или значимых данных в работе приложения на основе настроенных позитивных моделей приложения (совпадение с моделью или наоборот – отклонение от неё).
- Обнаружение аномалий и значимых параметров непосредственно внутри вложенных данных, передаваемых по протоколу HTTP без ограничений на количество уровней вложенности.
- Возможность обнаружения аномалий, свидетельствующих о возможных попытках атак, осуществляемых «методом грубой силы» (bruteforce).

– Подавление ложных срабатываний:

- Наличие функции упрощенного («быстрого») подавления ложных срабатываний оператором SolidWall AI Security Gateway непосредственно при просмотре описания выявленной аномалии.
- Возможность тонкой настройки подавления ложных срабатываний различных механизмов определения аномалий в привязке к отдельным параметрам запроса/ответа или бизнес-действиям.

– Принятие решений:

- Управление правилами принятия решений на основе данных об источнике (IP-адрес, пользователь, ID сессии) и цели HTTP-транзакции (приложение, бизнес-действие), а также обнаруженных в ней аномалиях или значимых данных.
- Поддерживаются следующие возможные решения: блокировать HTTP-транзакцию, пропустить HTTP-транзакцию, пометить HTTP-транзакцию, модифицировать ответ, добавить значение источника в список (черный, белый или серый, в зависимости от заданной политики).
- Управление правилами принятия решений (создание, удаление, перегруппировка) с помощью графического конфигуратора через интерфейс управления.

– Мониторинг и аудит событий информационной безопасности:

- Выводится обобщенная статистика по следующим параметрам: статусы HTTP-ответа сервера, текущей пропускной способности (входящий/исходящий трафик), задержки обработки запросов приложением, уровня враждебности среды.
- Сообщения об обнаруженных и/или заблокированных атаках на приложения выводятся в агрегированном виде, агрегация происходит по приложениям, типам обнаруженных аномалий и времени возникновения аномалий.
- Для событий безопасности агрегируется информация об источниках атак, в том числе по географическому распределению источников по странам, по номерам автономных систем.

- Возможность автоматического оповещения пользователей системы о событиях безопасности с использованием электронной почты, гибкая настройка оповещений из Веб-интерфейса.
- Возможность настройки выгрузки данных по отдельным заданным логическим действиям.
- Возможность автоматической генерации отчетов в формате «*.PDF», содержащих статистику событий информационной безопасности за заданные промежутки времени, в том числе регулярные отчеты с периодичностью «раз в сутки», «раз в неделю», «раз в месяц».
- Возможность передачи данных в режиме реального времени в систему мониторинга и корреляции событий информационной безопасности (SIEM).

2.3 Пользователи системы

В SolidWall AI Security Gateway присутствуют встроенные группы, описанные в таблице 1.

Таблица 1. Пользователи системы

Группа	Назначение
Администратор	Все члены группы обладают неограниченными правами и полный доступ ко всем функциям веб-интерфейса
Аналитик	Все члены группы обладают ограниченными правами: – просмотр и редактирование данных в разделах «Обзор», «События», «Правила», «Приложения», за исключением безвозвратного удаления событий безопасности, и только в рамках приложений, к которым имеют доступ; – генерирование отчетов и настройка уведомлений для своей учетной записи
Оператор	Все члены группы имеют ограниченный интерфейс (отсутствует раздел «Настройка»). Обладают правом на просмотр информации, связанной с приложениями, без права редактирования

Администратором SolidWall AI Security Gateway является пользователь, в обязанности которого входят:

– поддержание работоспособности технических средств;

– планирование архитектуры SolidWall AI Security Gateway;

– установка (инсталляция), эксплуатация и поддержание функционирования SolidWall AI Security Gateway (в т. ч. планирование требуемых аппаратных ресурсов);

– мониторинг журналов;

– настройка профилей и политик SolidWall AI Security Gateway с учетом особенностей, защищаемых ИИ-приложений;

– взаимодействие с аналитиками SolidWall AI Security Gateway.

Аналитиком SolidWall AI Security Gateway является пользователь, в обязанности которого входят:

– настройка профилей и политик SolidWall AI Security Gateway с учетом особенностей, анализируемых ИИ-приложений;

– мониторинг и анализ событий безопасности;

– принятие решений о реагировании на зафиксированные события;

– проверка (тестирование) корректности функционирования SolidWall AI Security Gateway при внедрении новых версий, защищаемых ИИ-приложений;

– взаимодействие с подразделением, эксплуатирующим ИИ-приложение при анализе сообщений об ошибках (особенно в отношении ложных срабатываний);

– взаимодействие с Администратором SolidWall AI Security Gateway.

Роль «Оператор» назначается пользователю SolidWall AI Security Gateway, в обязанности которого входят:

– регулярный мониторинг состояния защищаемых приложений;

– взаимодействие с аналитиками и/или администратором в случае нештатных ситуаций.