

ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК

программного обеспечения «Луми 42 ЛЛМ ФВ»

Параметр	Значение
Наименование ПО	«Луми 42 ЛЛМ ФВ»
Правообладатель	ООО «ЛУМИ 42»
Версия ПО	4.2
Год	2026

Оглавление

1. Назначение и область применения.....	3
2. Сведения о версии и составе поставки	3
3. Сценарии применения.....	3
4. Функциональная архитектура	3
5. Основные функциональные возможности	4
6. Интеграция, API и подключение LLM-провайдеров	5
7. Механизмы безопасности LLM-взаимодействий	6
8. Обработка файлов и мультимодальных данных	6
9. Аудит, аналитика и отчетность	7
10. Администрирование, пользователи и тенанты.....	7
11. MCP и контроль инструментов AI-агентов	7
12. Эксплуатационные характеристики.....	7
13. Роли пользователей и права доступа	8

1. Назначение и область применения

Программное обеспечение «Луми 42 ЛЛМ ФВ» версии 4.2 предназначено для контроля и защиты взаимодействий корпоративных приложений, чат-порталов, ИИ-ассистентов, ИИ-агентов и пользователей с большими языковыми моделями. Продукт выступает промежуточным техническим слоем между прикладной системой и LLM-провайдером, проверяет входящие запросы, файлы, контекст, ответы модели и вызовы инструментов.

Основная задача ПО - снизить риски prompt injection, jailbreak-атак, утечки секретов и иной чувствительной информации, несанкционированного использования LLM-сервисов, неконтролируемых действий AI-агентов и отсутствия трассируемого аудита при внедрении ИИ в бизнес-процессы.

В интерфейсе и технических материалах могут использоваться обозначения LLMFW и LLM Firewall. В настоящем документе они относятся к одному и тому же программному обеспечению «Луми 42 ЛЛМ ФВ».

Ключевое назначение

Продукт обеспечивает централизованную точку применения политик безопасности к LLM-трафику: до вызова модели, после получения ответа, при обработке файлов и при вызове инструментов AI-агентов.

2. Сведения о версии и составе поставки

Параметр	Значение
Официальное наименование	«Луми 42 ЛЛМ ФВ»
Версия ПО	4.2
Правообладатель	ООО «ЛУМИ 42»

3. Сценарии применения

ПО ориентировано на организации, внедряющие ИИ в прикладные сервисы, внутренние порталы, чат-ботов, ассистентов, агентские контуры, процессы поддержки, аналитики и автоматизации.

Группа пользователей	Потребность / сценарий
Команды информационной безопасности	Контроль prompt injection, jailbreak, утечек данных, секретов, действий AI-агентов, журналов безопасности и инцидентов.
Разработчики и интеграторы	Подключение приложений через OpenAI-compatible API, proxy, slug-based маршруты, SDK, Chat Portal или MCP Gateway.
Владельцы корпоративных данных	Предотвращение передачи ПДн, секретов, токенов, документов с коммерческой тайной и иной чувствительной информации во внешние LLM.
Compliance и внутренний аудит	Получение трассировки, событий аудита, отчетов, PDF-выгрузок, журналов администраторских действий и доказуемой истории решений.
Эксплуатационные команды	On-premise-установка, обновления, резервное копирование, диагностика, мониторинг ресурсов и подключение провайдеров.
Бизнес-подразделения	Безопасное использование чат-портала и LLM-сервисов без доступа к административной настройке продукта.

Типовые сценарии применения: пилотное подключение чат-бота, защита от утечек в промптах, контроль AI-агента, аудит использования ИИ, безопасная обработка файлов, подключение внутренних или внешних LLM-провайдеров, внедрение MCP-контура с политиками и согласованиями.

4. Функциональная архитектура

«Луми 42 ЛЛМ ФВ» состоит из набора функциональных модулей, работающих совместно как шлюз безопасности LLM-трафика и административная платформа управления политиками. Компоненты могут размещаться в инфраструктуре заказчика при on-premise-развертывании версии 4.2.

Компонент	Функциональное назначение
API Gateway / LLM proxy	Прием запросов приложений, проверка аутентификации, определение сценария, применение политик, маршрутизация в LLM и постпроверка ответа.
Admin UI	Настройка провайдеров, сценариев, правил PIM/SDM, пользователей, тенантов, SIEM, MCP, лицензии и обновлений.
Chat Portal	Веб-интерфейс для конечных пользователей, работающий через защищенный канал LLMFW с pre-check и post-check.
Policy Engine	Принятие решений allow, warn, redact, block, require_approval и применение настроек сценария.
PIM	Обнаружение prompt injection, jailbreak, попыток извлечения системных инструкций и обхода политик.
SDM	Обнаружение, маскирование и блокировка ПДн, секретов, токенов, ключей и конфиденциальной информации.
File Processing	Проверка MIME-типа, размера, архивов, active content, OCR, ICAP, извлечение текста и анализ содержимого.
Audit / Logging / Analytics	Запись событий, журналов, метрик, отчетов, аналитики, расследований и выгрузок.
MCP Gateway	Безопасный контроль MCP-клиентов, серверов, инструментов, ресурсов, prompt-шаблонов, политик и согласований.

5. Основные функциональные возможности

Функциональный блок	Описание функциональных возможностей
Защита промптов	Выявление prompt injection, jailbreak, role hijack, извлечения системных инструкций, obfuscation/evasion и многоходовых атак. Поддерживаются правила, пороги, ML-слой, таксономия угроз и тестовая песочница.
Защита чувствительных данных	DLP/SDM-проверки входов, файлов и ответов; обнаружение ПДн, секретов, ключей, токенов, паролей, банковских реквизитов и конфиденциальной информации; режимы block, redact, warn и allow.
Сценарии безопасности	Создание профилей для ботов, ассистентов и приложений: провайдер, модель, системный промпт, PIM/SDM-пороги, файловый профиль, логирование, use_mode и bypass.
Подключение LLM-провайдеров	Поддержка openai-compatible, anthropic, google, gigachat, локальных OpenAI-compatible endpoints, SOCKS5-прокси, теста связности и списков моделей.
API-интеграция	OpenAI-compatible chat/completions и responses, embeddings, models discovery, slug-based и header-based маршруты, работа с API-ключами тенанта.
Обработка файлов	Профили файлов, лимиты, MIME-типы, архивы, активное содержимое, OCR, ICAP-антивирус, извлечение текста, хранилище файлов и связь с аудитом.
Чат-портал	Пользовательский веб-интерфейс с потоковыми ответами, проверкой входов до вызова LLM, постпроверкой ответов, загрузкой файлов, per-tenant настройками и предупреждениями.
Аудит и расследования	События безопасности, trace_id, причины решений, детекции, FP-метки, инциденты, таймлайн, raw payload, сведения о latency, cost и usage.
Аналитика и метрики	Overview, Semantic Map, Flow Map, Threat Graph, Texts, Prompt Explorer, KPI, топы сценариев, провайдеров, правил, IP и PDF-отчеты.
SIEM-интеграция	Отправка событий в Syslog, HTTP/HTTPS webhook, Telegram; поддержка CEF/JSON, фильтров риска и тестовых сообщений.
Мультитенантность	Изоляция организаций, API-ключей, сценариев, пользователей, бюджетов, лимитов, логов и audit events по tenant_id.
MCP-контур	MCP Library и Secure MCP Gateway: клиенты, серверы, инструменты, ресурсы, prompts, политики allow/block/approval, billing, approvals и runtime logs.

6. Интеграция, API и подключение LLM-провайдеров

Для новых интеграций рекомендуется slug-based формат, при котором идентификатор сценария входит в URL. Для совместимости сохраняется header-based формат с заголовками x-api-key и x-scenario-id.

Маршрут / механизм	Назначение
POST /v1/<scenario_id>/chat/completions	OpenAI-compatible вызов чата через конкретный сценарий безопасности.
POST /v1/<scenario_id>/responses	Native responses flow для совместимых клиентов, потоковой передачи и расширенных возможностей.
POST /v1/<scenario_id>/embeddings	Получение embeddings через привязанный или fallback-провайдер сценария.
GET /v1/<scenario_id>/models	Получение списка моделей провайдера или каталога доступных моделей.
POST /v1/chat/completions + headers	Header-based совместимость со старыми интеграциями.
SOCKS5 proxy	Маршрутизация обращений к LLM-провайдеру через корпоративный или внешний туннель.
Chat Portal	Подключение пользователей к LLM через защищенный веб-интерфейс без прямой работы с API.
MCP runtime	JSON-RPC методы initialize, tools/list, tools/call, resources/list, resources/read, prompts/list, prompts/get.

При обработке запроса ПО проверяет API-ключ, определяет тенант и сценарий, загружает правила и лимиты, анализирует входящий текст и файлы, определяет провайдера и модель, отправляет запрос в LLM, проверяет ответ, записывает аудит, логи и метрики.

Примеры curl-команд, slug-based и header-based маршрутов приведены в руководстве пользователя и инструкции по интеграции. В настоящем документе они зафиксированы как функциональные интерфейсы продукта.

7. Механизмы безопасности LLM-взаимодействий

Механизм	Функциональное описание
PIM: лексический слой	Словари и регекс-паттерны на русском и английском языках для быстрого выявления инструкций обхода и опасных маркеров.
PIM: структурный слой	Анализ ролей, форматных разделителей, Markdown/JSON/YAML, zero-width символов, bidi overrides, confusables и иных техник обфускации.
PIM: ML-слой	Контекстная оценка сложных атак; настройки paranoia и weight позволяют управлять порогом и вкладом ML в итоговый риск.
Таксономия PIM	INSTR_OVERRIDE, ROLE_HIJACK, DATA_EXFIL, SYS_PROMPT_EXTRACT, SOCIAL_ENG, MULTI_TURN, ENCODING_EVASION, CONTEXT_OVERFLOW, JAILBREAK.
SDM: PII	ФИО, телефоны, email, паспортные данные, СНИЛС, ИНН, карты, счета, БИК, адреса, зарплаты, SMS-коды и пароли.
SDM: Secrets	API-ключи, токены, connection strings, приватные ключи, пароли, credentials и иные секреты.
SDM: Confidential	Конфиденциальные документы, ДСП, NDA, коммерческая тайна, кастомные паттерны и ключевые слова.
Режимы действий	allow, warn, redact, block; для MCP также require_approval. Для ответов LLM могут применяться отдельные output-политики.

Решение принимается до передачи данных провайдеру. Если выбрано redact, чувствительные фрагменты заменяются масками, например [ПДн], [ТЕЛЕФОН], [КАРТА], [SECRET], [ML_PII_1]. Если выбрано block, запрос останавливается до provider call.

8. Обработка файлов и мультимодальных данных

Файлы и изображения обрабатываются до вызова LLM. Профиль файлов задает лимиты, разрешенные MIME-типы, правила архивов, active content, OCR, антивирусную проверку, извлечение текста и срок хранения.

Этап конвейера	Назначение
Preflight	Проверка количества файлов, размера, MIME-типа по magic bytes и признаков архивных бомб.
Malware	Антивирусное сканирование через один или несколько ICAP endpoint-ов, если это включено в профиле.
Extraction	Извлечение текста из PDF, DOCX, PPTX, XLSX, TXT/CSV и других поддерживаемых форматов с лимитами страниц, листов и символов.
Canonicalization	Нормализация Unicode, BOM, пробелов и представления текста перед анализом.
OCR	Распознавание текста с изображений PNG, JPEG, GIF, WebP с учетом минимального порога уверенности.
Chunking	Разбиение крупного текста на source items, проходящие PIM и SDM-анализ.
Storage	Опциональное хранение файлов с retention_days, статистикой, фильтрами и ссылками на аудит по trace_id.

Типовые угрозы через файлы: prompt injection в метаданных, текстовые инструкции на изображениях, макросы, JavaScript в PDF, zip-бомбы, двойные расширения и фальшивые MIME-типы.

9. Аудит, аналитика и отчетность

Раздел	Функциональные характеристики
Логирование	Журнал входных и выходных сообщений: Original Input, Sanitized Input, Sent To Provider, Provider Output, Sanitized Output, Shown In Portal, фильтры и детальная панель.
Журнал системных событий	Авторизация, изменение пользователей, тенантов, сценариев, провайдеров, лицензий, обновлений, бэкапов и аварийных режимов.
Аудит	События безопасности с decision, risk_score, reasons, findings, masking, full I/O, usage, cost, files, portal metadata и raw payload.
Инциденты	Создание инцидента по trace_id, статусы open / in_progress / resolved / closed, комментарии и таймлайн расследования.
Алерты	Статистика срабатываний правил, hits, FP, last seen, анализ качества и тюнинг правил.
Аналитика 2.0	KPI, распределения решений, графы, карты, тексты, enrichment pipeline, drill-down в Audit и Logging.
Отчеты	PDF-отчеты аудита и аналитики, повторная загрузка, удаление, включение KPI и таблиц событий.
SIEM	Syslog, HTTP/HTTPS webhook, Telegram, CEF/JSON, фильтры риска и блокировок.

Логирование отвечает на вопрос «что прошло через тракт обработки», а аудит - «почему было принято конкретное решение политики». Для расследований используется общий идентификатор trace_id.

10. Администрирование, пользователи и тенанты

Функция администрирования	Описание
Тенанты	Логическая изоляция организаций, команд или проектов; привязка сценариев, провайдеров, пользователей, ключей, лимитов и событий.
API-ключи	Создание, хранение, описание назначения, ротация и отзыв ключей для gateway-интеграций.
Лимиты и бюджеты	RPM, burst, RPD, дневной и месячный бюджет, режим warn или block при превышении.
Пользователи	Создание учетных записей, роли, статус, пароль, granular permissions для операторов.

Функция администрирования	Описание
LDAP / AD	Интеграция с корпоративным каталогом, group mapping, portal user groups, strict group membership.
SAML 2.0 / OIDC	SSO, discovery, callback, attribute/claim mapping, роли admin/operator/user.
2FA	TOTP для усиления входа администраторов и пользователей.
Система	Лицензия, обновления, мониторинг CPU/RAM/диска, аварийный Global Bypass, диагностика и telemetry.
Резервные копии	Создание и восстановление бэкапов базы, конфигурации, сценариев, правил, пользователей и настроек.

11. MCP и контроль инструментов AI-агентов

MCP-раздел объединяет Secure MCP Gateway и MCP Library. Он предназначен для безопасного управления инструментами, ресурсами и prompt-шаблонами, доступными AI-агентам через Model Context Protocol.

Блок	Функциональность
MCP Library	Каталог версионизируемых пакетов: policy_pack, scenario_pack, server_profile, risk_taxonomy, forensic_template, rollout_recipe, bundle.
Жизненный цикл пакетов	Черновик, проверен, одобрен, выпущен, архив; копирование заводских пакетов и выпуск пользовательских версий.
Применение и shadow	Предпросмотр бандла и применение в режиме наблюдения без реальных блокировок для оценки влияния политик.
Манифест и дрейф	Снимок инструментов MCP-сервера, обнаружение новых, измененных и удаленных tools, одобрение или подавление дрейфа.
Раскатка	Canary, пилот тенанта, продакшен; gate-условия по блокировкам и нагрузке approvals, возможность отката.
Secure MCP Gateway	Клиенты, серверы, ресурсы, prompts, tools, policies, approvals, logs, billing, rate limits и secret check.
Runtime-решения	allow, require_approval, block, shadow; ошибки client_secret_invalid, tool_not_allowed, approval_required, mcp_quota_exceeded и др.

12. Эксплуатационные характеристики

Версия 4.2 может разворачиваться on-premise на сервере заказчика с использованием Ubuntu 22.04/24.04 и Docker. Поставочный архив устанавливается скриптом install.sh, который загружает Docker-образы из архива и запускает сервисы.

Параметр	Минимально	Рекомендуется
Операционная система	Ubuntu 22.04/24.04	Ubuntu 22.04/24.04 LTS с актуальными обновлениями безопасности
CPU	4 CPU	6 CPU и выше
RAM	8 GB	16 GB и выше
Диск	50 GB	100 GB SSD и выше
Контейнерная среда	Docker Engine и Docker Compose Plugin	Docker с настроенным автозапуском и резервным копированием
Сетевой доступ	HTTPS к порталу и API, исходящий доступ к LLM-провайдерам или локальным inference-сервисам	TLS, домен/внутренний адрес, правила firewall/proxy, мониторинг

Первичная настройка включает активацию лицензии, вход в админ-панель, смену пароля, добавление LLM-провайдера, создание сценария, API-ключа и тестовый запрос. Дополнительная задержка зависит

от состава включенных проверок, объема контекста, файлов, выбранной модели, сети и инфраструктуры размещения.

13. Роли пользователей и права доступа

Роль	Назначение	Права доступа
Super Admin	Технический суперпользователь и управление мультитенантностью.	Полный доступ, включая тенанты и глобальные настройки.
Admin	Администратор системы или организации.	Полный доступ к сценариям, провайдерам, PIM/SDM, аудиту, пользователям, системе.
Tenant Admin	Администратор конкретного тенанта.	Управление пользователями, сценариями и данными своего тенанта в рамках изоляции.
Operator	Оператор, SOC-аналитик, DevOps или аудитор с настраиваемыми полномочиями.	Гранулярный доступ full / read / none по разделам.
User	Конечный пользователь чат-портала.	Доступ только к Chat Portal и профилю без админ-панели.