Методы и подходы проектирования приватной AI-обработки чувствительных данных

Садова Карина
руководитель направления AI,
технологический предприниматель

9 лет

занимаюсь нейронками, начинала с их разработки

200+

моделей на проде

5 лет

управления командами и продуктами

✓ Читаю лекции
✓ Пробую запускать стартапы
✓ Консультирую бизнесы по вопросам разработки и интеграции AI в бизнес процессы

Садова Карина
руководитель направления AI,
технологический предприниматель

Мы научились быстро запускать ИИ-фичи

но часто забываем про безопасность

Садова Карина
руководитель направления AI,
технологический предприниматель

⚡ Скорость внедрения ≠ защита данных

- 🔷 Готовые API «из коробки» ≠ безопасны по умолчанию
- 🔷 Данные и промпты часто передаются без аудита и шифрования
- 🔷 Security-by-design остаётся опцией, а не стандартом

Садова Карина
руководитель направления AI,
технологический предприниматель

⚡ Некоторые из рисков

- API-ключи в открытом доступе → **утечка доступа и бюджета**
- Промпты с персональными данными → **хранение в логах провайдера**
- Отсутствие шифрования → **перехват данных в транзите**
- Prompt injection → **выполнение вредоносных инструкций**
- Нет трассируемости решений → **невозможность аудита и объяснения**
- Обработка данных вне юрисдикции → **штрафы по GDPR / 152-ФЗ**

Садова Карина
руководитель направления AI,
технологический предприниматель

OpenClaw

512 уязвимостей
8 критических
63% инстансов содержат уязвимости

25 000+ инстансов развернулись на открытом порте без аутентификации

12 812 инстансов позволяют удаленное выполнение кода одним кликом

7% навыков из ClawHub содержат инструкции, передающие чувствительные данные через контекст модели

Садова Карина
руководитель направления AI,
технологический предприниматель

SKILL.MD и маркетплейсы скиллов

Из 31 000+ скиллов:

- 26.1% содержат потенциальные уязвимости
- 5.2% имеют паттерны, указывающие на злонамеренное поведение

Садова Карина
руководитель направления AI,
технологический предприниматель

Проблема не в фреймворках, моделях или корпорациях,

а в бездумном их использовании as is

Садова Карина
руководитель направления AI,
технологический предприниматель

Риски при использовании API облачных моделей

- Хранение промптов — ваши запросы логируются и идут на тренировку моделей
- Утечка API-ключа → доступ к аккаунту, данным и бюджету
- Перехват трафика — если нет TLS или end-to-end шифрования
- Юрисдикция данных — обработка в стране с другими законами о приватности
- Небезопасный вывод — модель может вернуть вредоносный код, фейки, bias
- Зависимость от политик и доступности провайдера

Садова Карина
руководитель направления AI,
технологический предприниматель

Архитектуры - это спектр:
от cloud API до полностью local

Садова Карина
руководитель направления AI,
технологический предприниматель

Уровень	Архитектура	Плюсы	Минусы
Cloud API + Proxy	Публичные модели + ваш LLM Gateway и сервисы вокруг них	Быстро, дёшево, лучшие модели	Данные уходят наружу, vendor lock-in
Private Cloud / VPC	Azure OpenAI / AWS Bedrock в изолированном VPC, zero-retention policy	Комплаенс, приватная сеть, аудит	Всё ещё third-party модель, дороже
Hybrid Edge+Cloud	Локальная модель маскирует PII, а облако используется для сложного	Баланс: меньше утечек, экономия токенов	Сложнее архитектура, нужна edge-инфра
Full Local	Своя модель внутри периметра, full encryption stack	Макс. контроль, compliance-friendly	Инфра-затраты, нужна MLOps-экспертиза

Садова Карина
руководитель направления AI,
технологический предприниматель

Список компактных LLM моделей

- Llama 3.1 8B Instruct (GGUF, Q4_K_M) — ~6 ГБ VRAM
- Mistral 7B / Mixtral 8x7B — сильная логика, Apache 2.0
- Qwen 2.5 7B/14B — мультиязычная, код, reasoning
- Phi-3.5 Mini (3.8B) — работает на CPU, удивительно умна
- Gemma 2 9B — от Google, пермиссивная лицензия

Садова Карина
руководитель направления AI,
технологический предприниматель

Сценарий	Практики
Публичный API	🟢 PII-маскировка до отправки 🟢 API-ключи в secrets-manager (не в коде!) 🟢 LLM Gateway: прокси + валидация + логи 🟡 Rate-limiting + детекция prompt injection 🟡 Канареечные токены для теста на утечку
Private Cloud / VPC	🟡 Приватные эндпоинты (Private Link / VPC Peering) 🟡 Zero-retention в контракте (DPA/BAA) 🟡 Шифрование: TLS 1.3 (in transit) + AES-256 (at rest) 🟡 Фиксация региона обработки (data residency) 🔴 Аудит доступа через CloudTrail / Azure Monitor

Садова Карина
руководитель направления AI,
технологический предприниматель

Сценарий	Практики
Гибрид	🟡 Локальная модель для PII-обработки (Phi-3, Llama 8B) 🟡 Роутинг: sensitive → local, general → cloud 🟡 Кэширование частых запросов на edge 🔴 Тест качества локальной модели vs cloud 🔴 Fallback-логика при низкой уверенности
Локальная модель	🟡 mTLS между всеми внутренними сервисами 🟡 Vault/KMS для ключей + авто-ротация 🟡 Sandbox для исполнения кода (gVisor / Firecracker) 🔴 Confidential Computing (SGX/SEV/Nitro) 🔴 Immutable storage для логов + полный аудит
Универсальные практики	🟢 Аудит по OWASP Top 10 for LLM 🟢 Логирование запросов/ответов (с маскировкой PII) 🟢 Playbook на инцидент: отозвать ключи, изолировать, уведомить 🟡 Red-team тесты на injection / exfiltration 🟡 Ежеквартальный review зависимостей и политик вендоров

Садова Карина
руководитель направления AI,
технологический предприниматель

Confidential Computing / TEE

Защита данных не только «на полке» и «в пути» — но и «в работе»

Садова Карина
руководитель направления AI,
технологический предприниматель

Традиционная защита	Confidential Computing
🟢 At rest: AES-256 на диске	🟢 In use: данные зашифрованы даже в оперативной памяти
🟢 In transit: TLS 1.3 в сети	🟢 In use: даже админ облака не видит ваши данные
🔴 В момент обработки: данные в памяти в открытом виде	✅ В момент обработки: данные в зашифрованном энклаве (TEE)

Но есть нюанс:

- Производительность: оверхед 10-30% (зависит от workload)
- Сложность разработки: нужен спец. SDK, отладка усложняется
- Поддержка моделей: не все фреймворки работают «из коробки»

Early Production — доступно в AWS/Azure/GCP, но требует экспертизы

Садова Карина
руководитель направления AI,
технологический предприниматель

Сценарий 1

Сервис телемедицины. Пациент не хочет, чтобы его история болезней лежала в логах облачного провайдера.

Confidential Computing workflow:

- Клиент проверяет аттестацию энклава: действительно ли это тот сервис, за который он себя выдаёт?
- Если да — симптомы шифруются публичным ключом энклава и отправляются в облако
- Облачный сервер получает зашифрованный пакет, но не может его прочитать. Он лишь передаёт его внутрь энклава
- Внутри TEE данные расшифровываются, модель делает прогноз, результат снова шифруется и уходит пациенту

Результат: облачный провайдер оказал услугу, получил оплату, но никогда не видел конфиденциальные данные. Это compliance-friendly архитектура для медицины, финтеха, юриспруденции.

Садова Карина
руководитель направления AI,
технологический предприниматель

Сценарий 2

Проблема: данные нужны для обучения, но делиться ими нельзя

При federated Learning остается риск: что если центральный сервер, который агрегирует градиенты, скомпрометирован?

Confidential Computing закрывает эту брешь:

- Каждый участник шифрует свои градиенты перед отправкой.
- Агрегация происходит внутри энклава, где данные временно расшифровываются только для математической операции.
- Даже если злоумышленник получит доступ к серверу — он не сможет извлечь ни сырые градиенты, ни итоговую модель до момента рассылки.

Это позволяет создавать отраслевые консорциумы для ИИ, соблюдая GDPR, 152-ФЗ и коммерческую тайну одновременно

Садова Карина
руководитель направления AI,
технологический предприниматель

Докажи, что вычислил, не раскрывая данных

Садова Карина
руководитель направления AI,
технологический предприниматель

Zero-Knowledge пруфы

✅ Private Inference

Доказать, что модель обработала запрос корректно, не показывая сам запрос

✅ Model Integrity

Проверить, что использовалась именно заявленная модель (не подменена)

✅ Подтверждение свойств о данных пользователя

Вместо передачи самих данных пользователя

Садова Карина
руководитель направления AI,
технологический предприниматель

Ограничения

- Вычислительная сложность
  ZK-proof для GPT-2 = часы вычислений
- Размер доказательств
  мегабайты на каждый запрос
- Сложность разработки
  нужны криптографы в команде

Садова Карина
руководитель направления AI,
технологический предприниматель

Задача	Что доказываем	Сложность схемы	Время генерации	Статус
Баланс > 1 млн	Знание числа, удовлетворяющего условию	🟢 Низкая (несколько операций)	< 1 сек (смартфон/ноутбук)	🟢 Production (Zcash, Identity)
Инференс GPT-2	Корректное выполнение нейросети	🔴 Экстремальная (117 млн параметров → млрд операций)	Часы (нужны GPU-кластеры)	🔴 Research / Pilot
Инференс TinyML	Маленькая модель (например, классификатор)	🟡 Средняя (тысячи параметров)	Секунды/Минуты	🟡 Early Production

Садова Карина
руководитель направления AI,
технологический предприниматель

Почему данные становятся чувствительными?

Садова Карина
руководитель направления AI,
технологический предприниматель

Уровень	Причина чувствительности	Примеры данных
👤 Личный / Субъективный	• Приватность, стыд, репутация • Нежелание быть узнанным / классифицированным • Эмоциональная уязвимость	• Переписки, дневники, голосовые • Фото, биометрия лица/голоса • Поиск: «симптомы болезни», «как уволиться» • Финансовые привычки, долги
🏢 Коммерческий / Бизнес	• Конкурентное преимущество • Ноу-хау, интеллектуальная собственность • Риск репутационных/финансовых потерь	• Исходный код, архитектура систем • Стратегии, дорожные карты, pricing • Базы клиентов, контракты, NDA • Промпты и fine-tuning датасеты для ИИ

Садова Карина
руководитель направления AI,
технологический предприниматель

Коммерческий / Бизнес	• Конкурентное преимущество • Ноу-хау, интеллектуальная собственность • Риск репутационных/финансовых потерь	• Исходный код, архитектура систем • Стратегии, дорожные карты, pricing • Базы клиентов, контракты, NDA • Промпты и fine-tuning датасеты для ИИ
Юридический / Регуляторный	• Требования GDPR, 152-ФЗ, HIPAA, CCPA • Обязанность защиты по закону • Штрафы, иски, отзыв лицензии	• Паспортные данные, ИНН, СНИЛС • Медицинские записи, диагнозы • Биометрия, генетические данные • Данные несовершеннолетних
Социальный / Этический	• Риск дискриминации, стигматизации • Манипуляция поведением, микротаргетинг • Угроза демократическим процессам	• Расовая, этническая, религиозная принадлежность • Политические взгляды, членство в организациях • Ориентация, гендерная идентичность • Психологические профили, уязвимости

Садова Карина
руководитель направления AI,
технологический предприниматель

Безопасность — не пункт назначения, а направление

- 100% безопасности не существует — есть управление рисками
- Каждый трейдофф — осознанный выбор
  - скорость ↔ защита
  - стоимость ↔ устойчивость
  - удобство ↔ контроль
- Технический долг в безопасности имеет сложный процент

Садова Карина
руководитель направления AI,
технологический предприниматель

Увидимся тут:

телеграм

блог

линкедин

Спасибо и пока!

Садова Карина
руководитель направления AI,
технологический предприниматель