

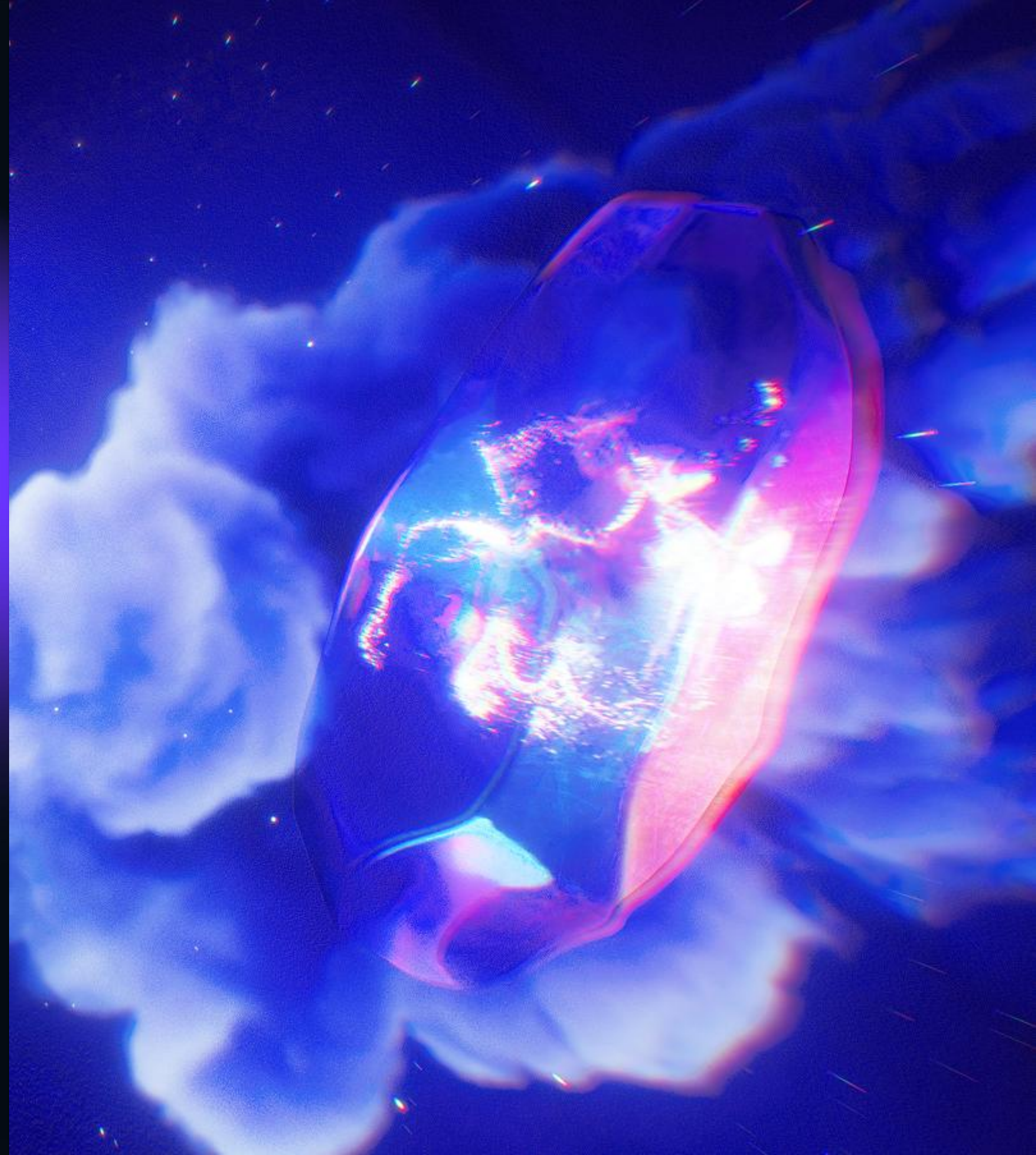
F6

ИИ в кибербезопасности



Станислав
Гончаров

Бизнес-руководитель департамента
защиты от цифровых рисков



Использование ИИ в Digital Risk Protection F6

F6

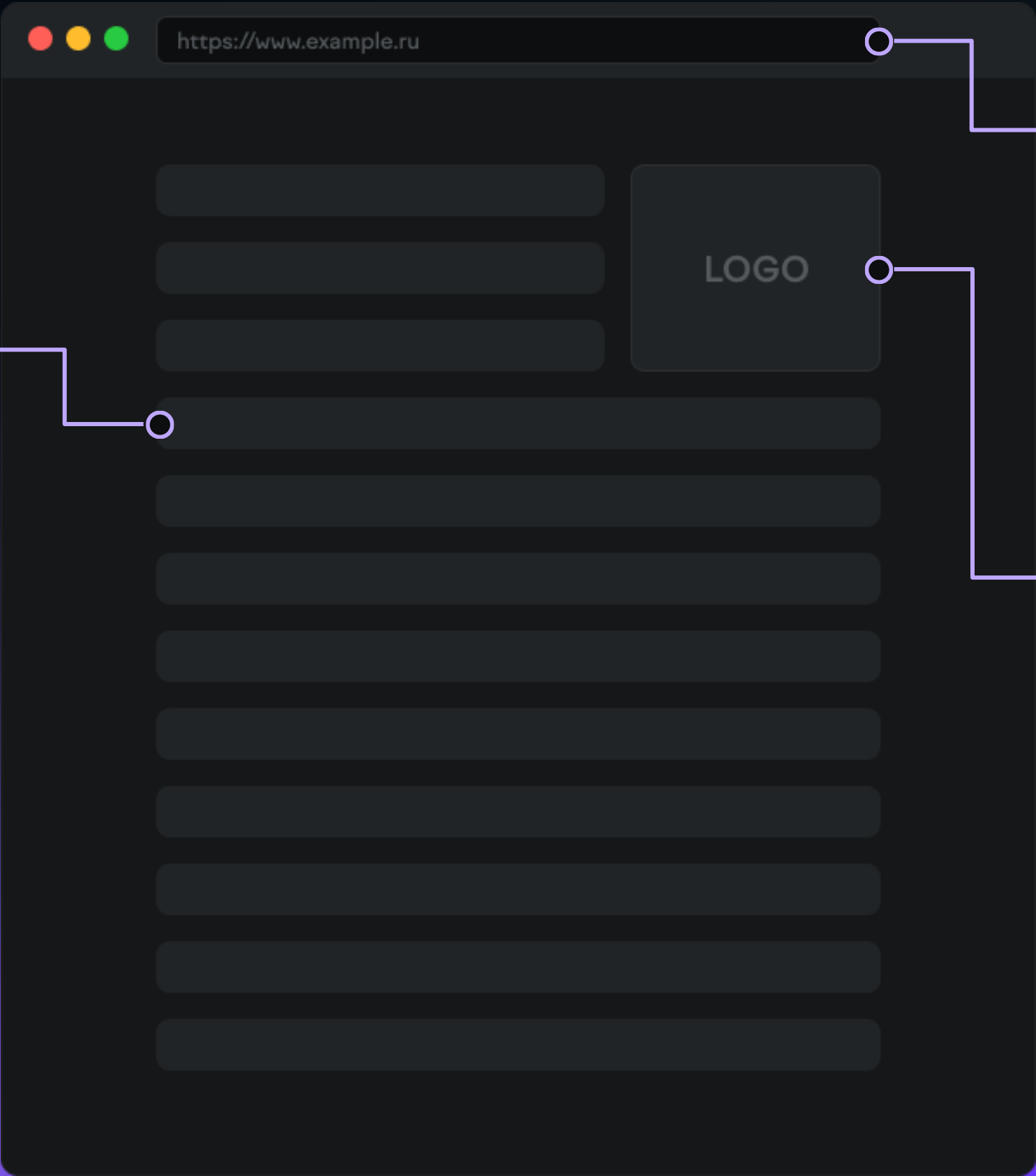
TCOBRA (Text-Content Oriented Brand Risk Analysis)

Анализ текстового контента веб-страниц по 5 метрикам для выявления потенциальных угроз по контексту.

Фрагмент 1

Фишинг 100% Скам 35%

Подарок Верификация Подтвердите, что вы не робот Введите номер телефона для быстрой верификации Продолжить Отмена Введите номер телефона Мы отправим вам код подтверждения ru +7 Введите корректный номер Продолжить Выберите страну ru Россия +7 kz Казахстан +7 by Беларусь +375 ua Украина +380 uz Узбекистан +998 kg Киргизия +996 am Армения +374 az Азербайджан +994 ge Грузия +995 tj Таджикистан +992 md Молдова +373 tr Турция +90 Введите код Мы отправили код подтверждения на ваш номер Неверный код Подтвердить Запросить новый код через 60 Где найти код? Введите пароль Пароль двухэтапной аутентификации MAX Неверный пароль Продолжить Готово! Вы успешно авторизованы Начать заново Где найти код? Для проверки на робота. Мы требуем проверки во избежание получения награды роботами. Понятно



UrlScan

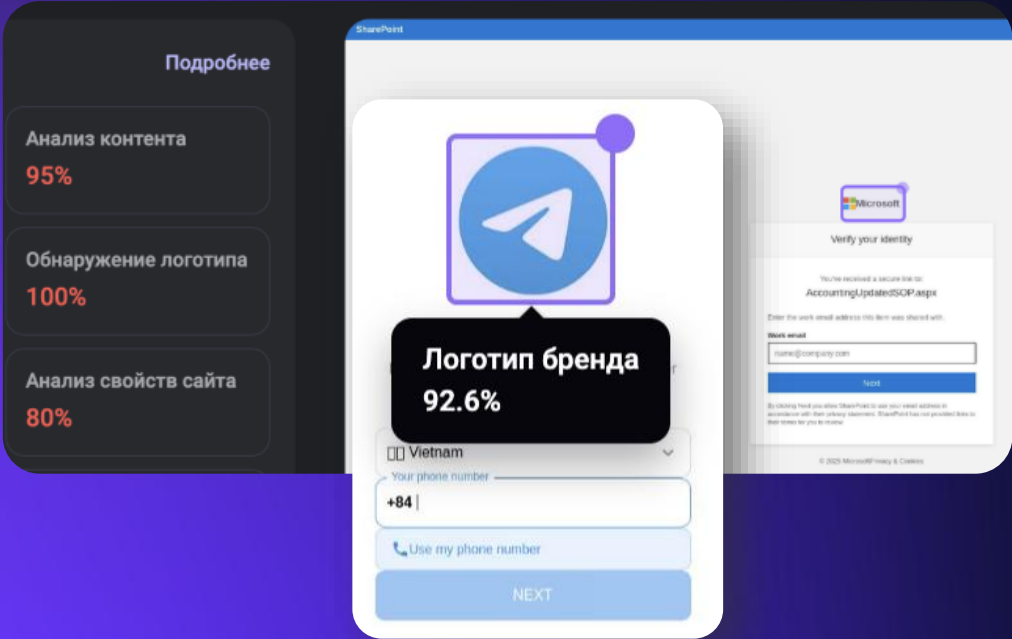
Анализ адресов сайтов для вероятностной оценки мошенничества.

Анализ домена

65%

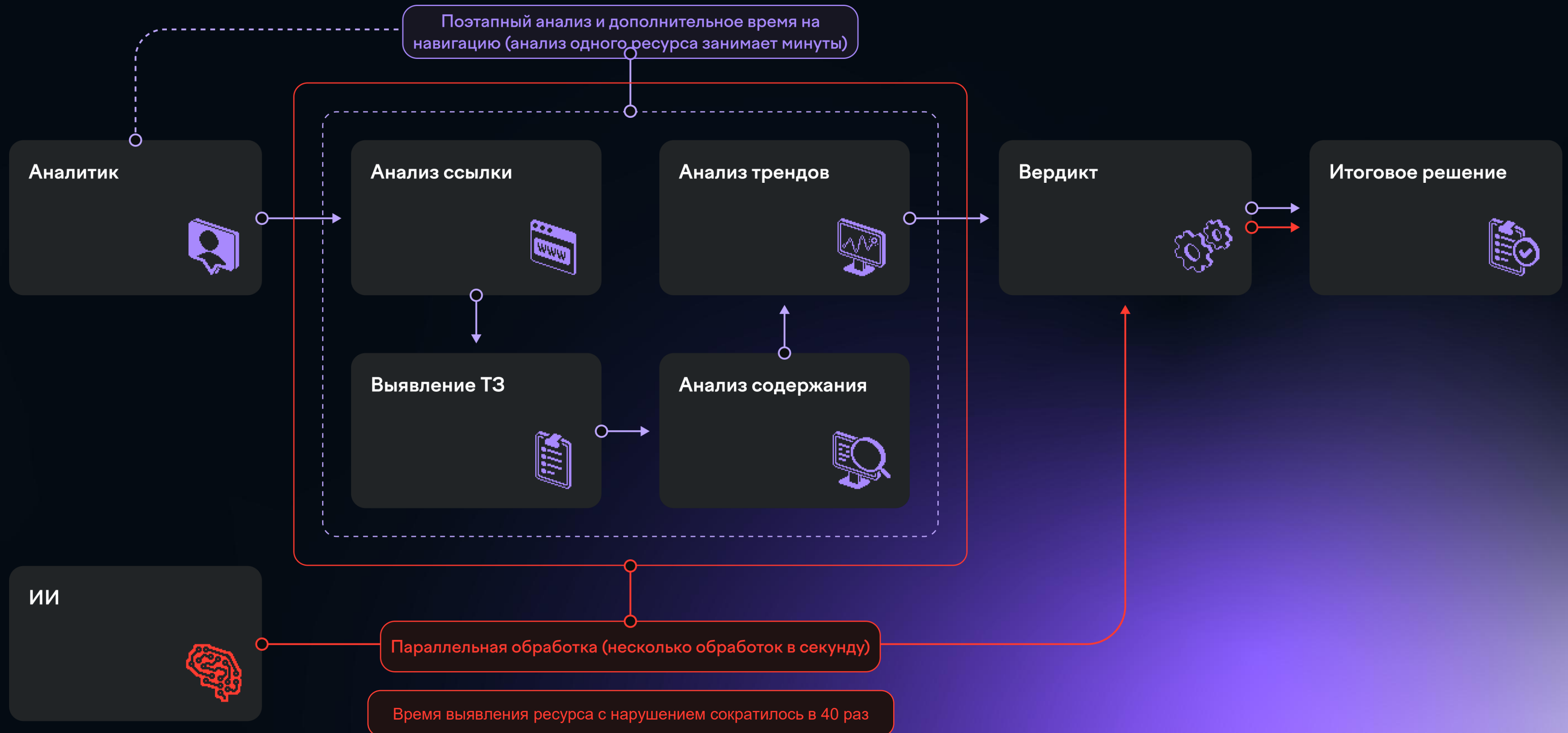
DeepVision

Анализ скриншотов веб-страниц для обнаружения логотипов.



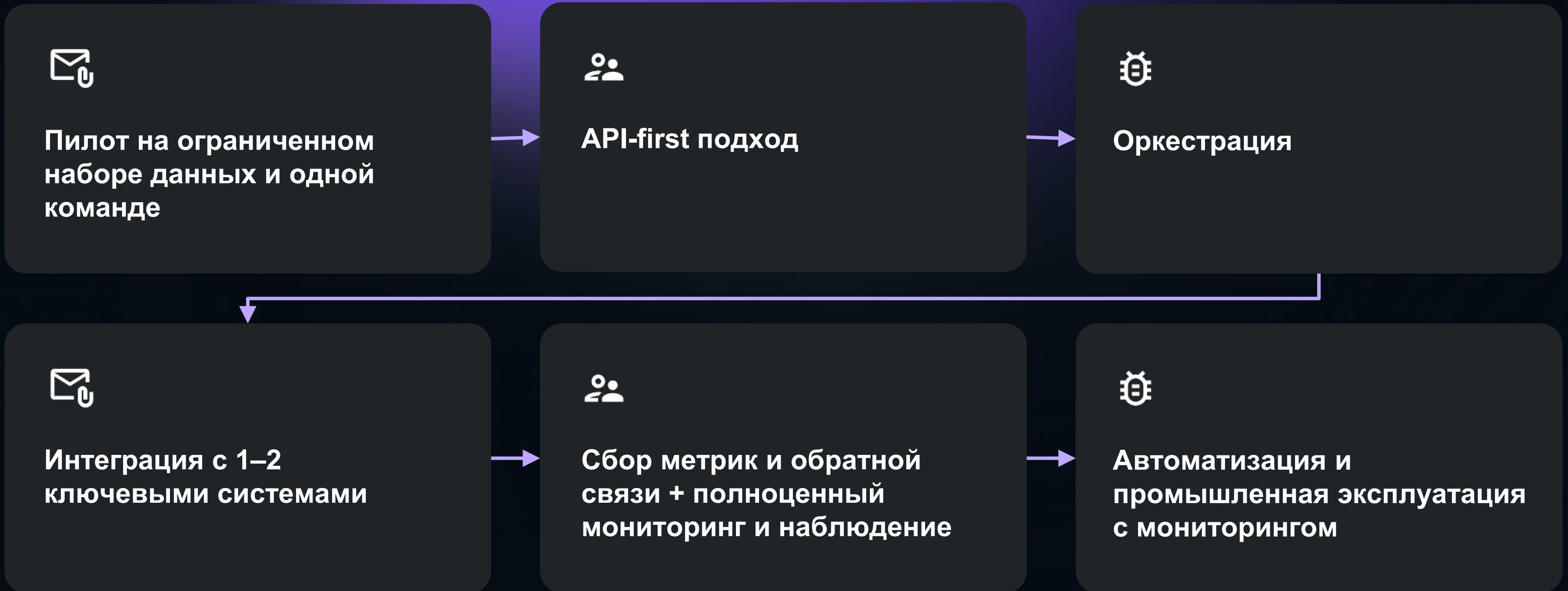
Оптимизация процессов и повышение эффективности

F6



Внедрение ИИ-решений

F6



Метрика	Норма (Низкий риск)	Аномалия / Срабатывание (Высокий риск)
Соотношение опасных и безопасных фрагментов	Преобладает чистый текст (без нарушений, без контента)	Высокая доля опасных категорий (ВПО, Фишинг, Скам)
Метрика переходов	Однородный контекст без резких скачков	Частая «зебра» (чередование опасных и безопасных зон)
Взвешенная унифицированная оценка	Низкая уверенность модели в угрозе	Высокие значения вероятностей от нейросети
Метрика по ключевым словам	Обычный текст	Совпадения с динамической базой ключевых слов
Метрика обнаружения брендов	Отсутствие защищаемых ТМ	Наличие брендов из кастомного списка клиентов



Внешние

- AI-генерированный фишинг и социальная инженерия
- AI-генерируемые фейковые видео и дипфейк аудио/видео
- Транслитерация и Unicode-обфускация
- Промпт-инъекции против ИИ-систем анализа



Внутренние

- Чрезвычайно высокое доверие к результатам ИИ
- Плохие промты
- Неоптимальные решения
- Завышенные ожидания



Генерация фишинговых и мошеннических страниц

С помощью промтов, задающих как визуальное оформление, так и динамику поведения сайта, злоумышленники способны за считанные минуты сформировать полностью готовый к публикации фишинговый сайт.



Использование deepfake

DeepFake-технологии позволяют воссоздавать образ публичных фигур и топ-менеджеров компаний, направляя трафик в мошеннические схемы: от рекламных роликов инвестиционных проектов до поддельных розыгрышей, размещаемых через взломанные аккаунты знаменитостей в социальных сетях.



Дезинформация с использованием генеративных моделей

Генеративные нейросети быстро создают компрометирующий и ложный контент, который, как правило, применяется в геополитически мотивированных кампаниях дезинформации.



Промпт-инъекции

Специально сформированные текстовые фрагменты внедряются в анализируемые файлы веб-ресурса, изменяя поведение LLM и позволяя обойти встроенные фильтры защиты.

Благодарим за внимание!

На связи 24/7

Технологии для борьбы
с киберугрозами

