



Дорога к безопасным ИИ-моделям

Инференс ИИ-модели $\neq$ продакшен-инференс

Технический директор,
основатель платформы «Боцман»
Михаил Кобик

Все хотят свой ИИ

Expectation: ИИ уже ждут в каждом продукте и процессе



АСТРА



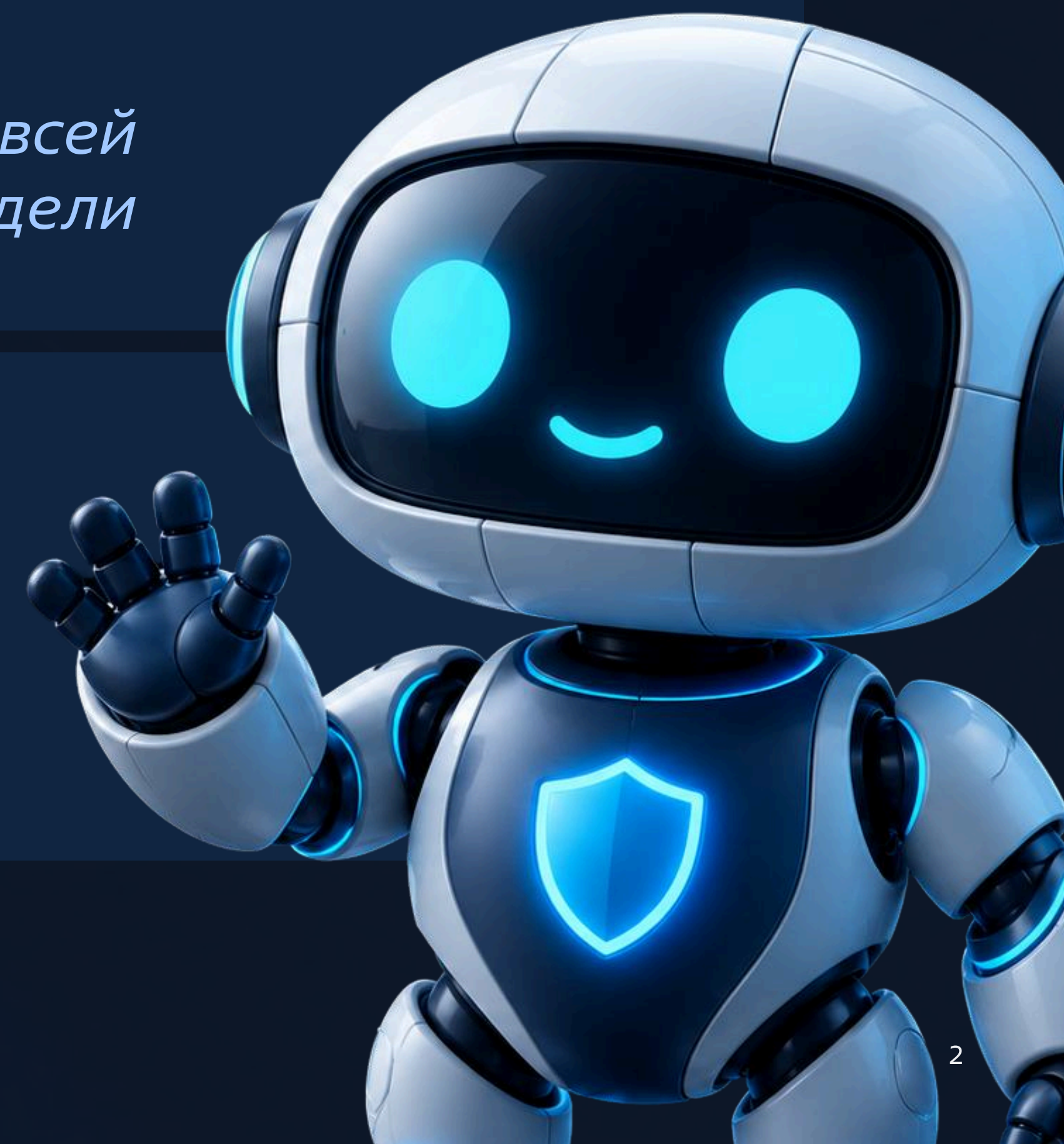
боцман

Свой ИИ — на своих данных, на своих серверах,
ПОД СВОИМ КОНТРОЛЕМ

*Спойлер: это свойство всей
системы вокруг модели*

Что значит
«безопасный ИИ»?

- Модель без багов?
- Данные без утечек?
- Сервис без простоев?
- Предсказуемые лимиты и доступ?



Популярная иллюзия: запустил модель — есть ИИ

Демо ещё не равно сервису



АСТРА



боцман

```
docker run --gpus all  
-p 8000:8000 \  
vllm/vllm-openai:v0.6.0 \  
--model meta-llama/Llama-3.1-8B-  
Instruct
```

```
curl http://localhost:8000/v1/chat/  
completions ...
```

Достаточно для:

- Демо
- Проверки качества
- Локальных экспериментов



Главный тезис

Безопасность появляется на уровне продакшен-сервиса



АСТРА



Боцман

Инференс

Модель отвечает



Продакшен-инференс

Модель работает как сервис

Надёжно

Масштабируемо

Безопасно

Наблюдаемо

Под контролем

Аналогия: двигатель ≠ автомобиль

Модель нужна, но безопасно «возит»
система вокруг неё



АСТРА



БОЦМАН

Модель — это двигатель.
Продакшен-сервис — автомобиль с тормозами,
ремнями и правилами

едет ≠ безопасно едет



Карта дороги: 5 граней безопасности ИИ-сервиса

Безопасный ИИ — это не одна кнопка,
а набор инженерных гарантий



АСТРА



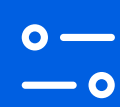
Боцман



Безопасный ИИ-сервис



Данные
дома



Управляемые
доступ
и лимиты



Контроль
утечки
критических
данных



Контент
проверен



Эффективный
кеш

Грань 1: данные остаются дома

Свой сейф, а не ячейка арендованная у чужого банка



АСТРА



боцман



Данные и модели «живут»
внутри периметра заказчика



Нет утечки по умолчанию,
которую пришлось бы
потом обосновывать

Грань 2: кто может «звать» модель

Банковская карта с лимитом и PIN, а не открытая касса



АСТРА



боцман

Доступ

- Единая авторизация
- Точечная выдача ресурсов



Контроль потребления

- Бюджеты
- Различные опции подписки



Грань 3: сервис нельзя использовать как вздумается



АСТРА



Боцман

«Такси довезёт» только до проверенных адресов из справочника

 Клиент



Боцман



 ИИ-модель

Грань 4: контроль того, что модель говорит

«Таможенный досмотр» на входе и на выходе



АСТРА



боцман



Запрос



Встроенные проверки



Модель

5 типов детекторов

Персональные данные

Секреты

Список запретов

Инъекции промтов

Произвольные конфигурации

Грань 5: недоступный ИИ = небезопасный ИИ

Надёжность — тоже безопасность



АСТРА



боцман

01

Управление
видеокартами

02

Высокая
доступность
(репликация)

03

Автоматическое
«исцеление»
моделей

04

Умное хранение
моделей



Если сервис «лежит» в критический момент — это уже инцидент

Дорожная карта: от GPU и модели до безопасного сервиса



АСТРА



боцман

Каждый шаг закрывает конкретную «дыру»

01

GPU + модель

02

Инференс «работает»

03

Масштабирование
модели

04

Авторизация

05

Биллинг
пользователей

06

Контроль
содержимого

07

Инфраструктура
вокруг модели

Безопасный
продакшен-инференс



Заказать демо «Боцман»

Михаил Кобик
Платформа «Боцман»
Технический директор, основатель

tg: @mkobik

mkobik@bootsman.tech

