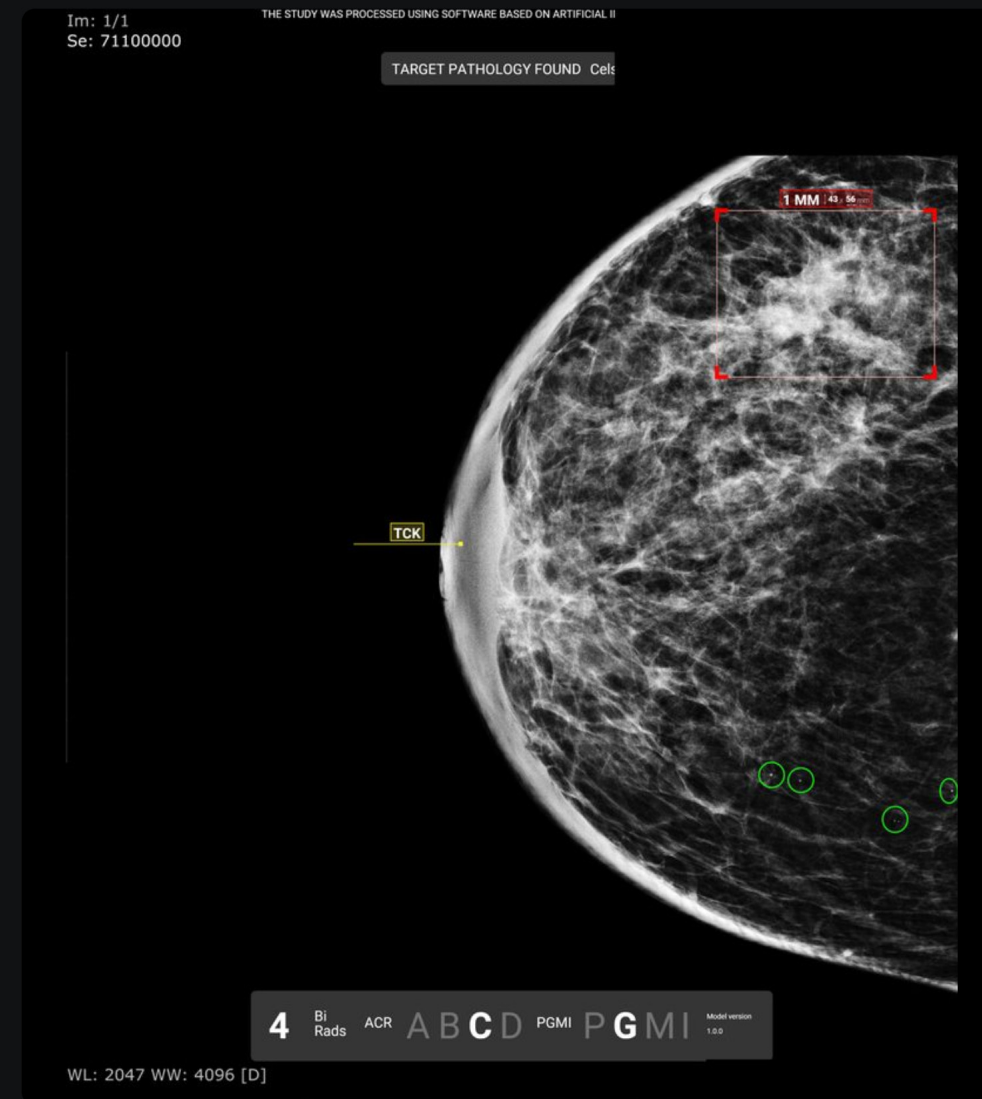


Как не пропустить рак?

Тестируем и мониторим медицинский ИИ:
от офлайн-метрик до LLM-дежурного

Женя Никитин СТО, Цельс



Кто я и что за Цельс

ИИ читает медицинские снимки: находит патологию, обводит её контуром и пишет заключение

7 лет

в продакшне,
зарегистрированное
медицинское изделие

6

модальностей: маммография,
флюорография и рентген лёгких, КТ
лёгких и мозга, МРТ малого таза

~700к

исследований в месяц

Я в Цельсе прошёл путь от единственного MLщика до СТО

Как это работает?



На каждое исследование отдаём картинку с обведёнными находками и текстовое заключение

СЦЕНАРИИ ПРИМЕНЕНИЯ

Второе чтение

врач смотрит сам, мы - вторая пара глаз

Сортировка рабочего списка

сначала те, где вероятность патологии выше

Автономная обработка

исследования без патологии врачу не показывают вообще

О чём будет доклад

1

ML-тесты

Клиническое качество
системы

2

Бэкенд-тесты

Принимает исследования и
формирует результат

3

Проверки в проде

Почему нельзя всё
проверить тестами?

4

ИИ-агенты

Что уже можно
делегировать?

Аппарат

PACS
больницы

Приём по
DICOM

Выбор серии

ИИ-модель

Рендер
результата

Отправка

PACS

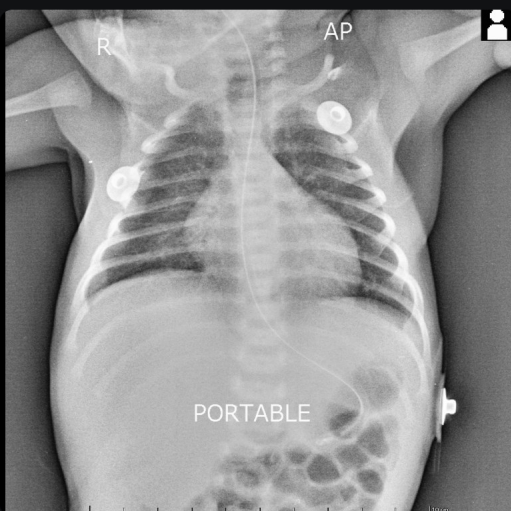
Вьюер врача

Внешняя оценка

Продакшн-среда

Мониторинг

Особенности входных данных



в сервис взрослой ФЛГ
прислали снимок
младенца, а модель на
таких не обучена

DICOM - теги одного снимка

в этом файле их 154

(0008,0060)	Modality	CR
(0008,0070)	Manufacturer	Shimadzu^Corp
(0008,1090)	Manufacturer's Model Name	DR-300
(0018,0015)	Body Part Examined	CHESTABDOMEN
(0028,0004)	Photometric Interpretation	MONOCHROME2
(0028,0010)	Rows	3032
(0028,0100)	Bits Allocated	16
(0028,1050)	Window Center	32768
(0028,1051)	Window Width	65536
(0008,0008)	Image Type	DERIVED, PRIMARY, ...

...и ещё 144

DICOM-файл: изображение плюс
куча мета-информации в тегах,
часть из них заполняет оператор



одно КТ - три серии: обзорный 2D, мягкотканная и
лёгочная реконструкции

Аппарат

PACS
больницыПриём по
DICOM

Выбор серии

ИИ-модель

Рендер
результата

Отправка

PACS

Вьюер врача

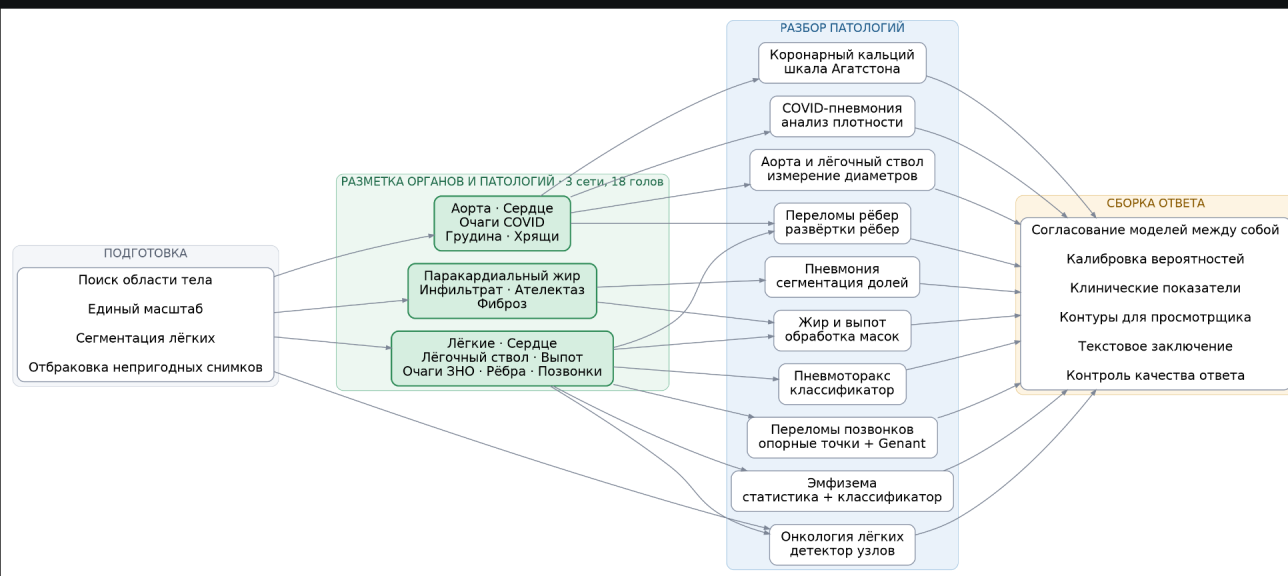
Внешняя оценка

Продакшн-среда

Мониторинг

ИИ-сервис - это не одна нейросеть

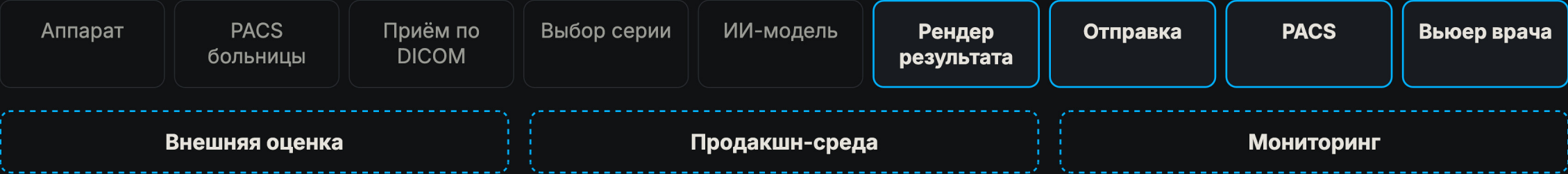
КТ ОГК - три сети, восемнадцать голов, шестнадцать независимых сервисов



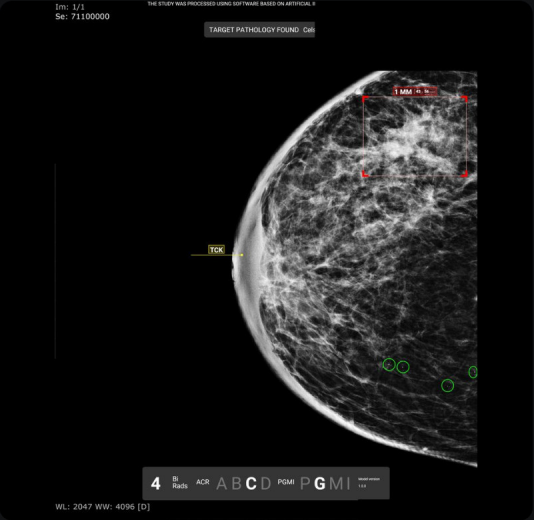
три сети на 18 голов размечают органы и патологии

поверх масок ещё десяток модулей считают клинические показатели: кальций, аорта, рёбра, пневмоторакс, онкология

16 независимых сервисов на GPU, очередь и общая память



Внешняя среда



врач открывает свой вьюер и видит картинку с вшитыми в пиксели находками и заключение

ОПИСАНИЕ ОЦЕНКИ ТРАКТОВКИ ПО КРИТЕРИЯМ				
	Полное соответствие	Нескорректная оценка	Ложноположительный	Ложноотрицательный
Рентгенографический признак	В заключении ИИ-сервиса описаны все целевые находки или указано «Целевая патология не выявлена»	Отмечено не менее 1 находки по целевой патологии (или не на всех снимках/проекциях)	Ложная/лишняя находка Описана ложная находка	Пропуск находки Не отмечено ни одной находки по целевой патологии, при фактическом их наличии
Злокачественное образование	В заключении ИИ-сервиса описаны целиком целевые находки. Верно выставлена категория BI-RADS (от BI-RADS 3 и выше). При отсутствии целевых находок указано	В заключении ИИ-сервиса описаны целиком целевые находки. Несоответствие категории BI-RADS фактической оценке, находящиеся в пределах BI-RADS 3 и выше	В заключении ИИ-сервиса описана целевая находка при её фактическом отсутствии, категория BI-RADS завышена	В заключении ИИ-сервиса не описана ни одна целевая находка при её фактическом наличии, категория BI-RADS занижена (1-

раз в месяц заказчик оценивает качество всех ИИ-сервисов по специальной методике



среда заказчика бывает любой - маммограф в кузове или медпоезд «Святой Пантелеймон»

1. ML-тесты

2. Бэкенд-тесты

3. Проверки в проде

4. ИИ-агенты

ЧТО ЗДЕСЬ ПРОВЕРЯЕМ

01

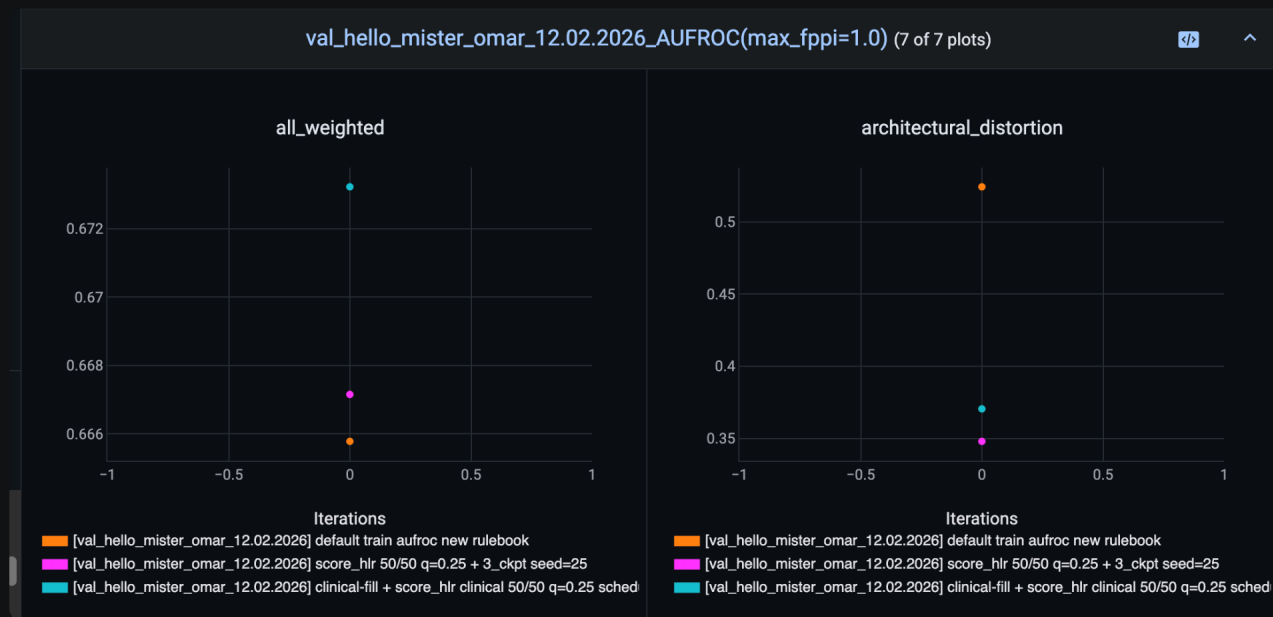
Клинические метрики модели

02

Корректность работы ML API

03

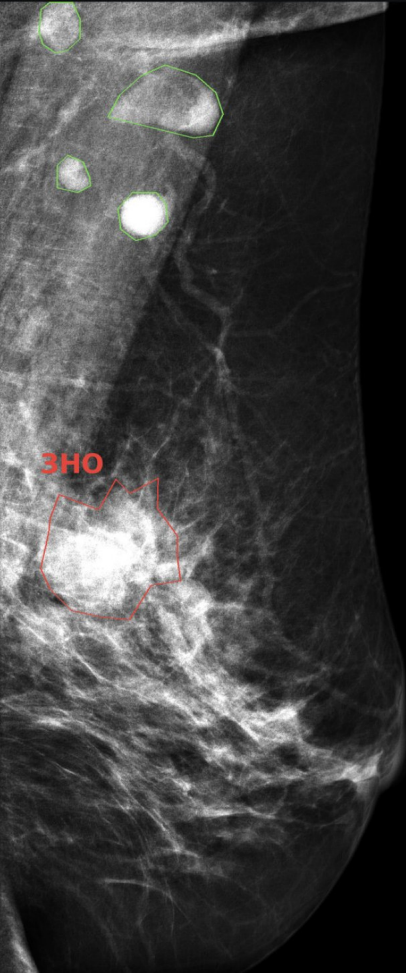
Скорость и надёжность
инференса



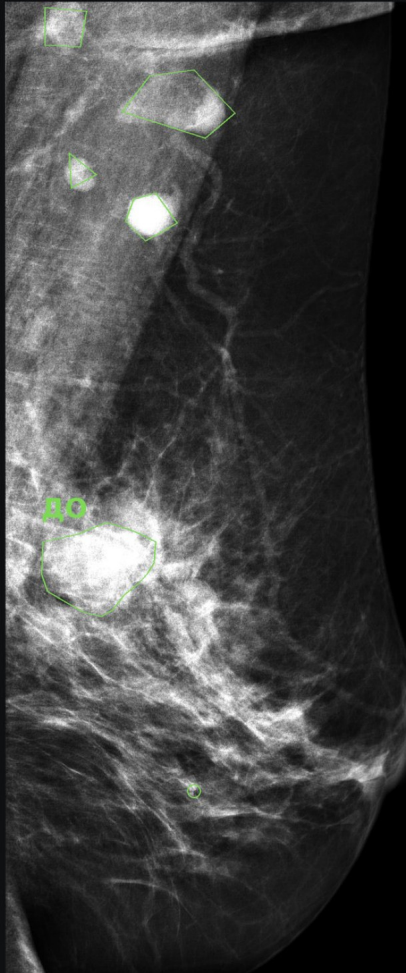
Сегодня поговорим про 01: как мы тестируем метрики модели, а не код ML-сервиса

три кандидата в ClearML на одном датасете: общая метрика отличается в третьем знаке, по одному классу - в полтора раза

врач А



врач В



54 %

расхождений между врачами
в оценке злокачественности

по 18 380 маммограммам с разметкой двух и более
врачей

Как измерять качество, когда нет оракула?

ПРАВИЛА И СУПЕРВИЗОР

Вносим изменения в правила разметки итерациями

Врач-супервизор следит за качеством разметки

КОНСИСТЕНТНОСТЬ РАЗМЕТКИ

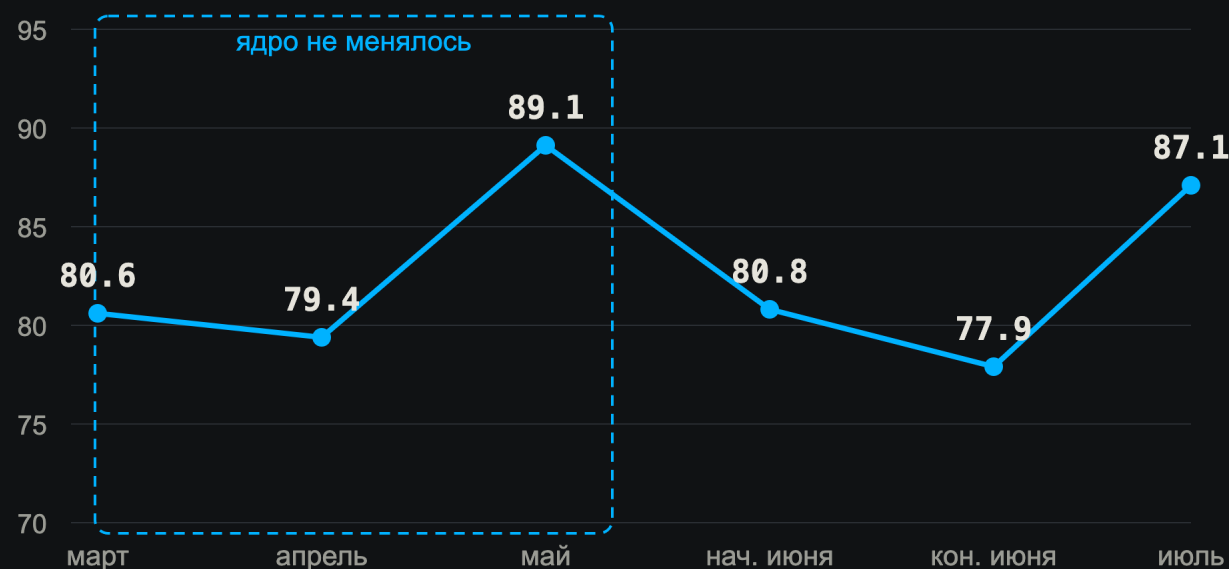
врача, который систематически размечает не как все, убираем из разметки - даже если в отдельных случаях он прав

ДИАПАЗОН ВМЕСТО ОДНОЙ ЦИФРЫ

метрику можно мерить на нескольких вариантах разметки: при широком разбросе вопросы к качеству данных

Три месяца на одной версии модели - разброс почти 10 пунктов

ВНЕШНЯЯ ОЦЕНКА ПО МЕСЯЦАМ, ОТ 0 ДО 100



ЧТО МОГЛО ПОВЛИЯТЬ

методику расчёта «тихо» поменяли

сменился врач, который оценивает

выборку собирают по нашим же предиктам

Это наша главная метрика - она должна быть выше порога и выше конкурентов.

Но **шума в ней часто больше, чем сигнала**: десять пунктов разброса без единого изменения в модели.

Поэтому нужна своя схема тестирования, которой можно доверять.

Процедура тестирования метрик - тоже продакшн-код

готовим релиз

прогон кандидата

вывод: не релизить

через месяц: модель
оказалась лучше

ЧТО БЫЛО НЕ ТАК

- забыли один из релизных датасетов
- пороги подобрали на старых весах
- прод и кандидата сравнили на разных версиях тест-сплита

ЧТО ДЕЛАЕМ

- у прогона своя версия: код, конфиг, данные
- тестируем код ML-тестов

Одна и та же модель может быть хуже или лучше другой в зависимости от

**источника
разметки**

**процедуры сбора
тестового датасета**

**багов в процедуре
тестирования**

1. ML-тесты

2. Бэкенд-тесты

3. Проверки в проде

4. ИИ-агенты

ЧТО ЗДЕСЬ ПРОВЕРЯЕМ

01

Корректность
приёма DICOM-
файлов и чтения
тегов

02

Правильность
валидации и отбора
серий

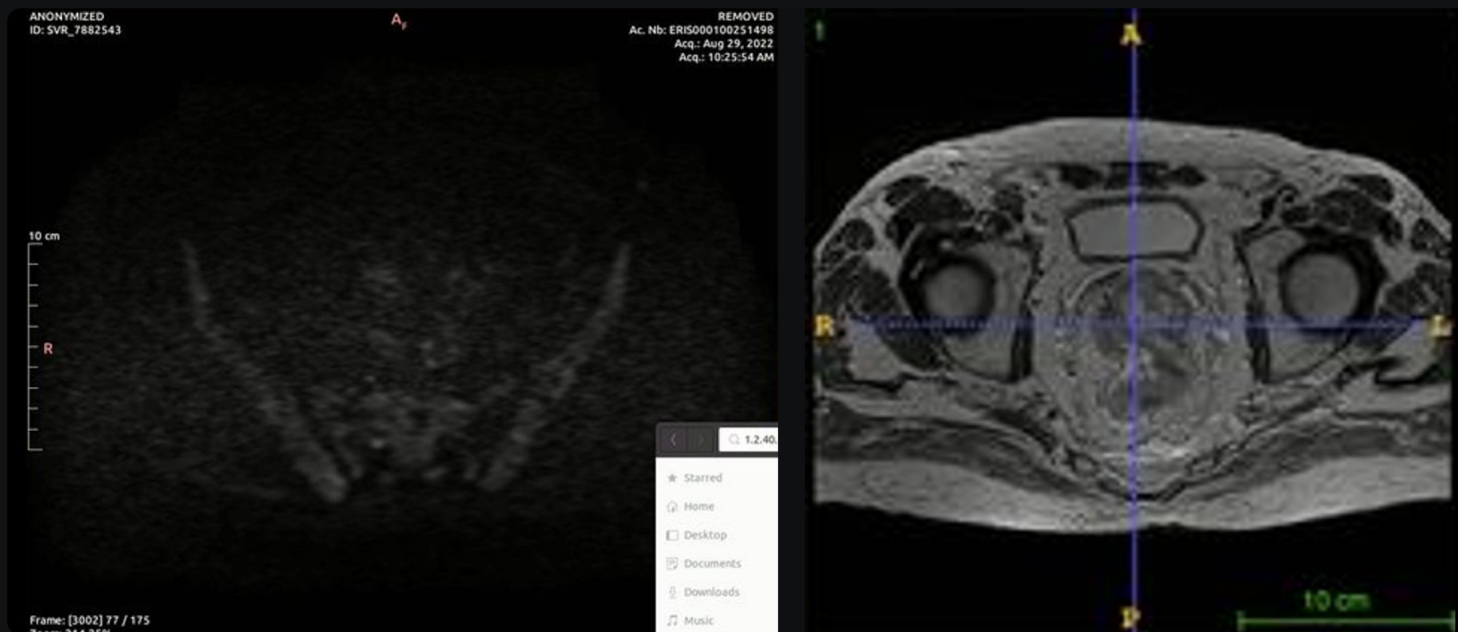
03

Корректность
картинки с
разметкой и
текстового
заключения

04

Доставка результата
в архив заказчика

Заказчик стёр два тега - метрика упала на четверть



слева - выбранная серия, справа - нужная

ЧТО СЛУЧИЛОСЬ

- при анонимизации стёрли два тега
- отбор серий молча взял другую серию
- ROC-AUC упал с 0.98 до 0.75

FAULT INJECTION

Убрали несущественный тег - результат не меняется.
Испортили критичный - явный отказ, а не тихая подмена серии

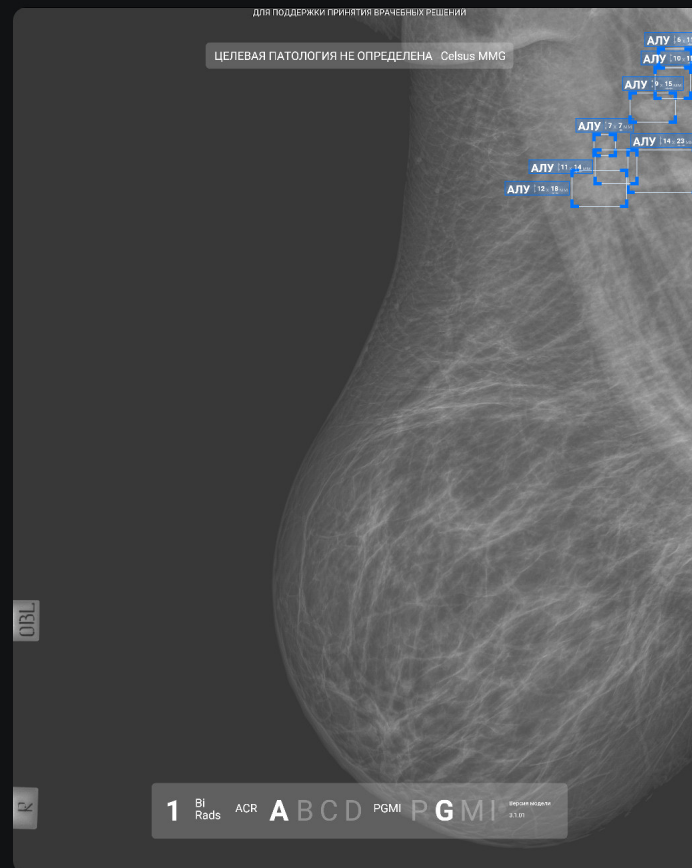
СТЁРТЫЕ ТЕГИ: БЫЛО → СТАЛО

(0008,103E) SeriesDescription	«AX T2»	→	тега нет
(0018,1030) ProtocolName	«AX T2»	→	«REMOVED»

Модель работает. Результат - некорректный



оригинал во вьюере



картинка с разметкой, которая уходит врачу

ЧТО СЛУЧИЛОСЬ

- в файле есть таблица яркости, но нет окна, и рендер её не читал
- предикт верный, смотреть неудобно

ЧТО ДЕЛАЕМ

- главный результат - картинка, поэтому без проверки глазами не обойтись
- дополнительные тесты на разницу с оригиналом

Регресс-пак: что в нём копится

артефакты съёмки

рамка · лишние объекты ·
надписи · поворот · «белая
грудь»

теги

FOR PROCESSING ·
Magnification · Modality* CR ·
нет PatientOrientation ·
инверсия

ошибки ядра и ложные срабатывания

то, что уже ломалось в
проде

разнообразие

Тверь · Дагестан · Брянск ·
Калуга

КОПИМ С 2021 ГОДА

10 наборов, почти 100 типов сложных
случаев

КОМБИНАТОРИКА ОСОБЕННОСТЕЙ

оцифрованная плёнка, белая рамка и кривой
PixelSpacing: каждый признак по
отдельности встречался, вместе - ни разу

Регресс-пак хорошо ловит то, что уже ломалось, но медицинские данные всегда удивляют

* Modality - тег с типом исследования: MG - маммография, CR - рентген, CT - КТ

Одна и та же модель даёт верный или неверный результат в зависимости от

**того, какую серию
ей отдали**

**того, как
нарисовали
картинку**

**сочетания
особенностей
снимка**

1. ML-тесты

2. Бэкенд-тесты

3. Проверки в проде

4. ИИ-агенты

ЧТО ЗДЕСЬ ПРОВЕРЯЕМ

01

**Совместимость с конкретным
PACS заказчика**

02

**Корректность работы под
реальной нагрузкой и пиками**

03

**Полнота обработки потока:
процент успешных обработок**

Все PACS* принимали, а один рвал соединение

наша картинка с разметкой

у пикселей картинки (PixelData)
нечётная длина в байтах

десятки PACS

молча дописывают нулевой байт

один строгий PACS

рвёт соединение

DICOM PS3.5 · 7.1.1 · VALUE FIELD

«...shall be of an even number of bytes in length»

Баг было невозможно найти, пока не попался привередливый PACS

* PACS (Picture Archiving and Communication System) – система хранения и передачи медицинских изображений

Не обрабатываем валидные исследования



УСЛОВИЕ В КОДЕ

```
BodyPartExamined.Contains("abdomen")
```

ЗНАЧЕНИЕ ТЕГА В ЭТОМ СНИМКЕ

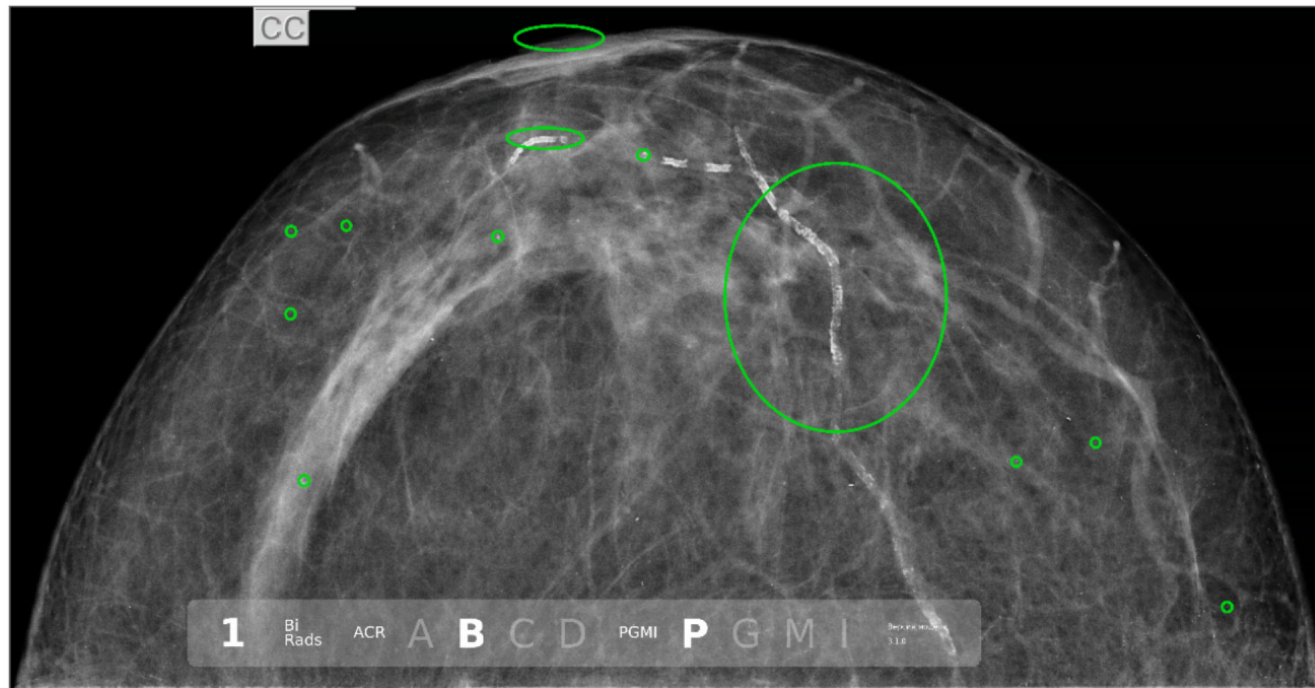
BodyPartExamined = CHESTABDOMEN

должно было отсекасть живот, а отсекло и грудь с животом

Постоянно нужно анализировать поток на новые случаи некорректной обработки тегов

Нестандартное железо

error in DeformablePSROIPoolForward: no kernel image is available for execution on the device



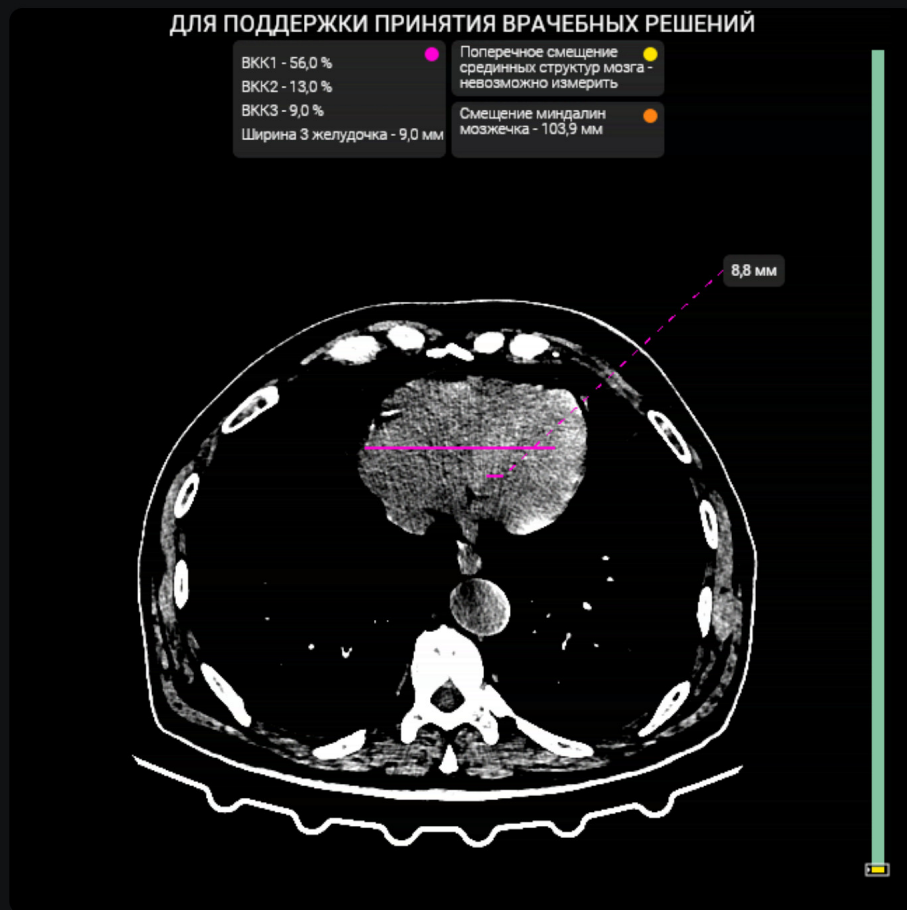
ЧТО СЛУЧИЛОСЬ

- карта 2016 года: у нас такой нет, до установки не прогнать
- модуль пишет ошибку, процесс жив, инференс возвращает нули, /health зелёный

ЧТО ДЕЛАЕМ

- self-test модулей сети при старте на целевом железе
- health check реальным инференсом

Хочется обрабатывать больше, но не любой ценой



ЧТО СЛУЧИЛОСЬ

- чтобы поднять процент обработки, ослабили фильтрацию по одному из тегов
- в сервис КТ головного мозга поехали другие части тела
- тесты зелёные, алертов нет: исследование обработано успешно

«ну мы нашли желудочки. правда, предсердий...»

ЧТО ДЕЛАЕМ

- ослабили фильтр на входе - проверяем, что ещё стало проходить
- периодический визуальный и автоматический контроль входных данных

**Анализ продакшна находит новые «тихие» ошибки
и пополняет набор тестов**

1. ML-тесты

2. Бэкенд-тесты

3. Проверки в проде

4. ИИ-агенты

ЧТО ЗДЕСЬ ДЕЛАЕМ

01

Собрать сырые
алерты в понятный
инцидент

02

Ответить, авария это
или шум

03

Разобрать то, на что
у людей не хватает
часов

04

Дать ответ тому, кто
спросил

Ловим больше, чем успеваем разбирать

облако

нет потока · недоставка
техдефекты · время
обработки

регионы

SSH до ноды · контейнеры
GPU · лаг очереди

федеральная платформа

по каждому региону:
нас отключили · перевели в
резерв · поток упал

проба

реальное исследование
каждые 10 минут

карточка инцидента INC-...

дежурный: авария, шум или ожидаемое?

Информации стало больше, чем времени у дежурных и аналитиков

Половину проблем в этом докладе нашли ИИ-агенты или люди с их помощью

CHESTABDOMEN в теге части тела

разбор отказов на потоке

рендер яркости: неверная кривая

сверка с референсной реализацией

всплеск находок после релиза маммографии

цифру с прода принёс бот в чате

новый заказчик, чужие снимки

далее в докладе

Сам мониторинг у нас тоже уже построен на LLM

алерт-менеджер собирает инциденты · ИИ-дежурный делает триаж · агент готовит недельный отчёт · бот отвечает на вопросы в чате

Алерт-менеджер: LLM собирает сырые алерты в один инцидент

✓ Инцидент #INC-E34053 • 47м • ЗАКРЫТ

Модальности: КТ ОГК

Клиенты: ЕМИАС Эксперимент

Начало: 2026-09-01 09:48:30 • Обновлён: 10:35:32 • Алертов: 4

Тип: 🕒 sla_timeout (6× прибито бэкендом по таймауту)

Состав: 6 прибито бэкендом по таймауту / 1 прочих

Поток: 19 за 40м (медиана 23, x0.8)

Вывод: Фундаментальная причина: рост суммарного потока на 30-40% за месяц (до 1.6-1.8к иссл./день) и пиковая нагрузка от пачек исследований в интервале 09:30-09:45. Рост трафика ЕМИАС в 3 раза создает дефицит ресурсов для платных потоков. Планируется введение приоритезации (Мосмед) и переконфигурация сервисов после обновления модуля позвоночника. ✎ Edited



↩ 40 replies

Following

ЧТО ДЕЛАЕТ LLM

- склеивает срабатывания в один инцидент: модальность, клиент, окно
- держит его открытым, пока авария идёт
- дописывает вывод по ходу разбора

ИИ-дежурный: триаж с уликами

 AlertBot BOT 2w ago

 Авто-триаж INC-E34053

 **Вердикт:** реальная авария (проверено)

Что: Облачное ML-ядро КТ ОГК (`celsus-core-kate-production-complex`) в насыщении — исследования копятся в очереди перед AI и прибиваются нашим SLA-таймаутом у обоих кастомеров КТ ОГК, не только у ЕМИАС.

Действие: Scale-out GPU-реплик ядра КТ ОГК (`detection / roiseg / lung-segmentation / postprocessor-coronary-calcium` , сейчас прибиты к 3 T4-нодам) с эскалацией в `ml_backend` · [needs-confirm]

Улики:

- `tracked_operations` (свой замер по 30-мин окнам, кастомеры 5+34): очередь «перед AI» p50 0.2→9.4с, p90 174с, сам AI p50 83→200с, 39 исследований с `AiProcessingStarted` без `Finished` ; «Загрузка» 23→36с — выросла, но

вердикт: реальная авария

 AlertBot BOT 24 hours ago

 Авто-триаж INC-49B398

 **Вердикт:** реальная деградация без подтверждённого импакта (проверено)

Что: `celsus-core-emias-flu/fluere-main` держит 2/4 реплики — оба пода на GPU-ноде `kube-worker-t4-production-gpu-fl-main-2` лежат с 05.09, поток ФЛГ ЕМИАС при этом идёт штатно.

Действие: удалить два залипших пода (`fluere-main-67d586bcd-mr2tr` , `-wgk2t`), чтобы ReplicaSet пересоздал их на `gpu-fl-main-2` · [needs-confirm]

Улики:

- k8s: оба пода в `Terminated / Exit Code 1` , Finished 05.09 08:29 и 09:02 UTC — деградация идёт 9 суток, а не «более 15 минут» из карточки (15m — это порог `for` правила, не длительность).
- Лог упавшего контейнера: `RuntimeError: can't start new thread` — исчерпание потоков/памяти на ноде-2 (два пода по лимиту 8Gi из нода 16Gi, allocated limits 112%); kubelet за

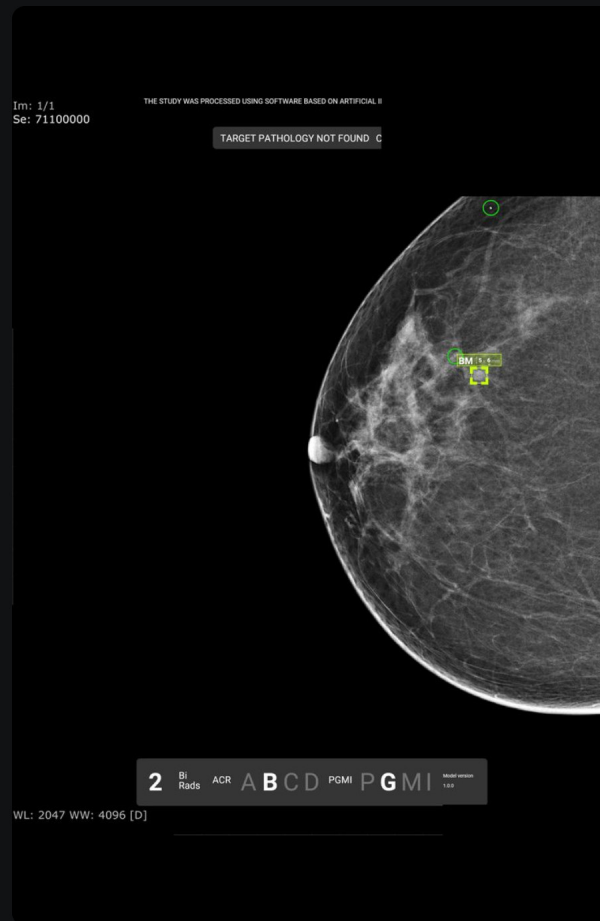
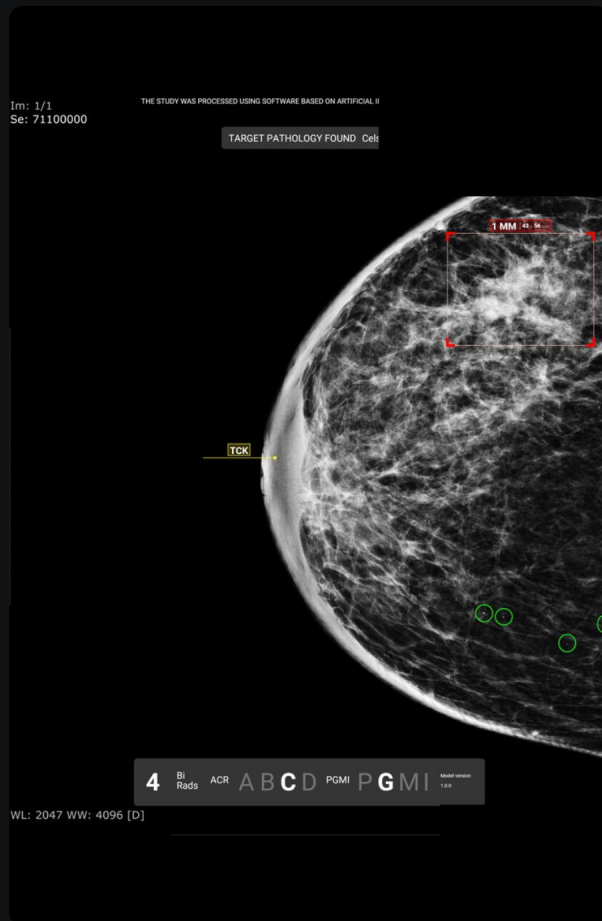
вердикт: деградация без импакта

ФОРМАТ ОТВЕТА

- вердикт
- что происходит
- предлагаемое действие
- улики
- чего не проверил

каждую улику можно
открыть и проверить руками

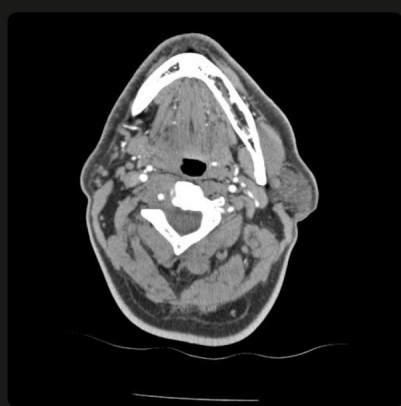
Агент проверяет картинки вместо части ручного тестирования



ЧТО ПРОВЕРИЛ АГЕНТ

- 20/20 обработаны
- находка на одной проекции видна на второй в той же зоне
- проекции распознаны верно, тексты отрисованы
- в пиксели вшиты ФИО, дата рождения, клиника

Еженедельный отчёт: агент разбирает отбраковку



КТ ОГК, Краснодарский край, «Недостаточная длина лёгких». В приоритетной серии — аксиальный срез шеи на уровне гортани и щитовидной железы. Лёгких в исследовании нет: отбраковка верна, вопрос в том, что регион прислал под кодом КТ ОГК.

priority · BodyPartExamined=HEAD · TOSHIBA · WC 40 / WW 450 · срез 0,5 мм



То же исследование, отброшенная серия: снова шея, срез 4 мм. Обе серии одного и того же анатомического региона, грудной клетки нет ни в одной.

excess · BodyPartExamined=HEAD · TOSHIBA · WC 40 / WW 350 · срез 4, мм

ЧТО СДЕЛАЛ АГЕНТ САМ, БЕЗ АЛЕРТА

- посчитал, сколько исследований мы не обработали и сколько это стоит
- пошёл в архив и открыл те, что мы отбраковали
- разобрал их глазами и объяснил каждое

ВЫВОД

отбраковка верна: грудной клетки нет ни в одной серии, регион шлёт КТ головы и шеи под кодом КТ лёгких

Каждый теперь - немного тестировщик

ПРОДЖЕКТ

 13:04
посмотри, по какой причине и корректно ли были отклонены следующие исследования:
1.2.9.2.6.1.192.168.0.11.16838.20260826104942.350
1.2.9.2.6.1.192.168.0.11.16838.20260826103207.2495
1.2.9.2.6.1.192.168.0.11.16838.20260825100837.1528
 6 replies | Follow

БОТ

Отклонены корректно: во всех трёх только боковая проекция, прямой нет

сверил логи приёма и отрендерил снимки; лид бэкенда ответила «Верно»

КООРДИНАТОР РЕГИОНОВ

 21:00
проверь, есть ли проблема с мосмедии - уже 30 минут ни одного исследования. есть ли проблема? если есть, то решаем ли мы ее?
 5 replies | Follow

БОТ

Проблема есть, но не на нашей стороне: МосМедИИ перестал направлять задачи всем вендорам

наш приём жив; через полчаса бот сам написал в тред «поток пошёл»

Раньше каждый такой вопрос был задачей на инженера: сходить в базу, выгрузить, объяснить

Когда данные доступны без посредника, вопросов становится больше, и всплывает то, на что раньше никто не смотрел

Who watches the watchmen?

У ПЛАГИНА АГЕНТА ЕСТЬ СВОИ ТЕСТЫ

кейсы

рабочий вопрос агенту
+ данные на тот момент
+ чем проверяем ответ

оценщики

модель-судья по критерию
маркер в ответе
вызвал ли нужный скилл

прогон без плагина

разница в оценке:
скилл вообще влияет?

Прогон без плагина показывает, даёт ли скилл хоть что-то

Что забрать с собой

В ML-системах баги почти всегда тихие

взяли в обработку мусор · отсеяли валидное исследование · модель выдала ерунду, а все алерты зелёные

Тестовый код тоже надо тестировать

у прогона свои версия, конфиг и баги, а по нему решают судьбу модели, которую делали месяцами

Метрику, которую считаете не вы, вы не контролируете

разметка, выборка и правила могут поменяться без вас

Прод - лучший источник новых тестов

регресс-пак помнит прошлое, новые случаи ищем на проде

ИИ сильно помогает тестировщику и дежурному

позволяет анализировать ранее недоступные объёмы информации

Заменит ли искусственный интеллект врача?

Частично уже да

Как и тестировщика

Спасибо за внимание!

Женя Никитин · СТО Цельс

t.me/varim_ml

