

Worksheet Выбора Версии YOLO

Ответьте на восемь вопросов ниже до выбора версии YOLO. Одна страница; один проект; одна печать.

1. Железо — где модель крутится в продакшене

- ☐ Облачный NVIDIA GPU (T4, L4, A10, A100, H100) — любая версия.
- ☐ NVIDIA Jetson edge (Orin Nano, AGX Orin) — v11 или v12.
- ☐ Apple Silicon (M-серия, iOS A-серия) — v11 лучше, без FlashAttn.
- ☐ Только Intel CPU / iGPU — v11 или YOLO26 (CPU-оптимизирована).
- ☐ Браузер (WebGL / WebGPU) — только v8 или v11.

2. Бюджет латентности на кадр

- ☐ До 5 мс — v10 (NMS-free экономит 4–6 мс) или v12 (FlashAttn).
- ☐ 5–20 мс — дефолт v11, v12 если требует точность.
- ☐ Более 20 мс — любая, выбор по точности / стоимости.

3. Пол точности (COCO-эквивалент mAP)

- ☐ Максимум на nano-размере (40%+) — v12-N.
- ☐ Edge-friendly с минимумом параметров — v9 или v11-N.
- ☐ Максимум на большой модели — v11-x или v12-x.

4. Лицензионный путь

- ☐ Закрытый коммерческий продукт — нужен Ultralytics Enterprise.
- ☐ Open-source продукт — AGPL-3.0 ок, исходники открываются.
- ☐ Внутренний инструментарий — уточнить score AGPL с юристом.
- ☐ Юр-согласование до экспорта (неделя 1, а не неделя 12).

5. План данных для обучения

- ☐ Кастомный датасет размечен (500–5 000 изображений на класс).
- ☐ Валидационный сет отдельно от тренировочного — без утечек.
- ☐ Дообучить 50–250 эпох от COCO-чекпойнта.
- ☐ Пайплайн Roboflow / CVAT / Label Studio выбран.

6. Deployment-рантайм + квантизация

- ☐ TensorRT INT8 — NVIDIA GPU, Jetson edge (дефолт для cloud).
- ☐ OpenVINO INT8 — Intel CPU и iGPU.
- ☐ CoreML FP16 — Apple Silicon (iOS / macOS / Vision Pro).
- ☐ TFLite INT8 — Android, edge-SoC, микроконтроллеры.
- ☐ ONNX — фолбэк / кросс-платформа.

7. Скоуп мульти-задач

- ☐ Только детекция — v9, v10, v11, v12 подходят.
- ☐ Детекция + сегментация + pose — v8 или v11 (5 задач).
- ☐ Oriented bounding box (повёрнутые объекты) — v8 или v11.
- ☐ Classification-голова — v8 или v11.

8. Стоимость миграции — только если меняете модель

- ☐ Бюджет на re-разметку / переекспорт оценён.
- ☐ Время дообучения на текущем датасете измерено.
- ☐ План A/B-теста: shadow новой модели рядом со старой 2 недели.
- ☐ Путь отката определён, если точность просядет.

Сводка из восьми чисел

Железо · бюджет латентности · пол точности · лицензия · план данных · deployment-рантайм · мульти-задачи · стоимость миграции.

Если хоть одно из восьми не закрыто, решение о версии YOLO будет принято по неправильной причине. Закройте все восемь до экспорта первой модели.