YOLO Version Selection Decision Worksheet

Answer the eight questions below before committing to a YOLO version. One page; one project; one print.

1. Hardware target — where the model runs in production

- ☐ Cloud NVIDIA GPU (T4, L4, A10, A100, H100) — any version.
- ☐ NVIDIA Jetson edge (Orin Nano, AGX Orin) — v11 or v12.
- ☐ Apple Silicon (M-series, iOS A-series) — v11 best, no FlashAttn.
- ☐ Intel CPU / iGPU only — v11 or YOLO26 (CPU-optimised).
- ☐ Browser (WebGL / WebGPU) — v8 or v11 only.

2. Latency budget per frame

- ☐ Under 5 ms — v10 (NMS-free saves 4 to 6 ms) or v12 (FlashAttn).
- ☐ 5 to 20 ms — v11 default, v12 if accuracy demands it.
- ☐ Over 20 ms — any version, pick on accuracy / cost.

3. Accuracy floor (COCO-equivalent mAP)

- ☐ Highest available at nano scale (40%+) — v12-N.
- ☐ Edge-friendly with smallest params — v9 or v11-N.
- ☐ Big-model maximum accuracy — v11-x or v12-x.

4. Licence path

- ☐ Closed-source commercial — Ultralytics Enterprise required.
- ☐ Open-source product — AGPL-3.0 acceptable, ship source.
- ☐ Internal-only tooling — confirm AGPL scope with legal.
- ☐ Pre-export legal sign-off complete (week 1, not week 12).

5. Training data plan

- ☐ Custom dataset labelled (500 to 5,000 images per class).
- ☐ Validation set separate from training set — no leakage.
- ☐ Fine-tune for 50 to 250 epochs from COCO checkpoint.
- ☐ Roboflow / CVAT / Label Studio pipeline picked.

6. Deployment runtime + quantisation

- ☐ TensorRT INT8 — NVIDIA GPU, Jetson edge (default for cloud).
- ☐ OpenVINO INT8 — Intel CPU and iGPU.
- ☐ CoreML FP16 — Apple Silicon (iOS / macOS / Vision Pro).
- ☐ TFLite INT8 — Android, edge SoCs, microcontrollers.
- ☐ ONNX — fallback / cross-platform.

7. Multi-task scope

- ☐ Detection only — v9, v10, v11, v12 all work.
- ☐ Detection + segmentation + pose — v8 or v11 (5 tasks).
- ☐ Oriented bounding box (rotated objects) — v8 or v11.
- ☐ Classification head — v8 or v11.

8. Migration cost — only if replacing an existing model

- ☐ Re-label / re-export budget estimated.
- ☐ Retrain time on existing dataset measured.
- ☐ A/B test plan: shadow new model alongside old for 2 weeks.
- ☐ Rollback path defined if accuracy drops.

The eight-number summary

Hardware target · latency budget · accuracy floor · licence path · training data plan · deployment runtime · multi-task scope · migration cost.

If any of these eight is unanswered, the YOLO version decision will be made for the wrong reason. Answer all eight before exporting the first model.