

Model Artifact + Procurement Audit Checklist

Companion to 'Model Artifact Formats and the Open-vs-Closed Procurement Decision'

Step 1 — Privacy & compliance filter

- ☐ Is any data the feature processes regulated (HIPAA, PCI, GDPR special category, defence)?
- ☐ Have you confirmed your data residency contract permits the API region you plan to call?
- ☐ Has legal reviewed the EU AI Act Article 50 transparency obligations applicable Aug 2, 2026?
- ☐ If self-hosting: do you have a documented model provenance and supply-chain scanning policy?

Step 2 — Capability gap filter

- ☐ Is the task frontier reasoning (GPQA-class) or production-grade (summarisation, ASR, CV)?
- ☐ Have you benchmarked at least one open-weight model on a sample of YOUR data?
- ☐ If using closed: do you have a documented fallback path if the vendor deprecates a model?
- ☐ Have you logged the licence (Apache 2.0 vs custom) of every open-weight you plan to ship?

Step 3 — Cost crossover

- ☐ Estimated tokens per day at launch _____; at 12 months _____.
- ☐ Closed-API cost per million tokens (input + output, blended) \$_____.
- ☐ GPU-hour cost in your deployment region (L4 / A100 / H100) \$_____.
- ☐ Expected concurrent batch size and throughput per GPU _____.
- ☐ Crossover volume (closed = open all-in) calculated and documented?

Step 4 — Engineering capacity tax

- ☐ Headcount available: ML engineer __, DevOps __, SRE oncall __, security __.
- ☐ Monthly fully-loaded engineering cost of the open-weights stack \$_____.
- ☐ Documented runbook for GPU shortage / region outage / vLLM crash?
- ☐ Documented model upgrade & A/B testing process for new open-weights releases?

Step 5 — Hybrid routing plan

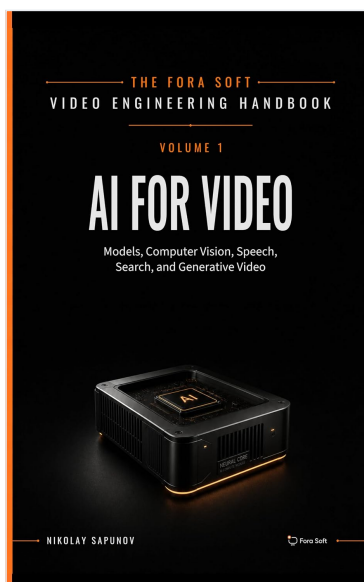
- ☐ Router rule defined for 'easy 80%' (open weights) vs 'hard 20%' (closed)?
- ☐ Per-request cost telemetry instrumented (closed tokens, open GPU-seconds)?
- ☐ Quarterly review on the calendar to re-run the cost arithmetic?
- ☐ Rollback plan if open-weights quality regresses on a model bump?

Format checklist for the artefact you ship

- ☐ Mobile app: model converted to Core ML (iOS) or LiteRT / ONNX (Android)? Not Safetensors?
- ☐ Browser feature: ONNX (WebGPU) or llama.cpp WASM (GGUF)? Bundle size measured?
- ☐ Edge server: ONNX Runtime + TensorRT EP, or vLLM Safetensors? Throughput measured?
- ☐ Cloud LLM serving: Safetensors via vLLM, or compiled TensorRT-LLM engine?
- ☐ If pickle-only: refused, or scanned with picklescan and quarantined?
- ☐ Every quantisation choice (Q4_K_M, Q5_K_M, Q8_0) backed by a quality benchmark?

Companion to: 'Model Artifact Formats — Open vs Closed Procurement'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/model-artifact-formats-procurement-decision



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This checklist walks the open-vs-closed decision through privacy, capability, cost crossover, team capacity and artifact format. Volume 6 gives the background behind each step: the AI feature lifecycle, deployment topologies, model artifacts and formats, and the ML minimum a product lead needs to decide.

WHAT YOU WILL FIND INSIDE

- Model artifacts: GGUF, ONNX, Core ML, Safetensors
- Deployment topologies: cloud API, edge, on-device
- The AI feature lifecycle from pilot to production
- ML basics for build-vs-buy decisions

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Choosing between a closed API and open weights for your video AI feature?

We will run your volumes through the cost crossover, benchmark an open model on your data and recommend a routing plan. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)