# Model Artifact + Procurement Audit Checklist

Fora Soft Learn

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 SaaS region you plan to call?
- [ ] Has legal reviewed the EU AI Act Article 50 transparency obligations for your specific use?
- [ ] 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 ___, on-call ___?
- [ ] 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?