@demotoprod · BUDGET · also LEVERS

Local or cloud checklist

A cloud API call hands your prompt to another party. For personal or regulated data, that brings in contract, retention and transfer rules. Five afternoon-sized checks settle what may go to a cloud model, what stays on a local one, and what each route costs per task.

1 Label each data class and write its route

Health-data guidance says a cloud provider that maintains protected health information for a covered entity is a business associate under HIPAA, even when it holds only encrypted data and lacks the key. FIPS 199 rates information LOW, MODERATE or HIGH. Each level gets a route.

- List every data class your AI reads (tickets, contracts, health records, code) and give each a level.
- Enforce one route per level in the gateway; unlabelled data gets the strictest.

```
low:      cloud
moderate: cloud, signed terms
high:     local only   # and any unlabelled data
```

2 Get retention and sub-processor terms in writing

A regulator's guide to processor contracts lists: act only on documented instructions, delete or return the data at the end, authorise any sub-processor. One API provider's guide says abuse logs are kept up to 30 days by default and zero retention needs prior approval.

- Request each provider's terms, tick those three items, record retention days and training use.
- A class whose terms fall short moves to the local route.

3 Pin the processing region as well as storage

A transfer has three cumulative criteria in the EU data-protection board's guidelines, one being an importer in a third country; remote access from there counts, even only displaying data on a screen, and so does storage in a cloud outside the EEA.

- Record where each route processes prompts and where logs and stored objects rest.
- For routes leaving your region, have your privacy lead confirm the transfer basis.

4 Keep a local fallback and test it on your prompts

Reliability guidance says a component should keep its core function when a dependency is unavailable, even degraded. A local model is a different model: in one study of 1B to 405B parameter models, smaller ones could suffer severe accuracy drops at 4-bit quantization.

- Run a sample of real prompts through both models and write down the gap.
- Fail over cloud to local only; drill it by blocking the cloud endpoint in staging.

```
if level == "high" or not cloud_up:
    return local    # never fail over to cloud
return cloud
```

5 Compute cost per task for both routes

One 2025 cost analysis finds on-premise viable mainly at 50M tokens a month or more, or under strict residency rules. Break-even typically runs from a few months for small models to 5 years for larger ones, counting hardware and electricity only.

- Log input and output tokens per task for a week; price each at its own per-million-token rate.
- Divide local hardware, power and staff time by monthly tasks; move a class only if local wins.

Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Paste into your project

ONE ROUTE PER DATA LEVEL, IN CONFIG

```
# routes.yaml: one route per data level
unlabelled: high
levels:
  low:      {route: cloud}
  moderate: {route: cloud, requires: [signed_terms, pinned_region]}
  high:     {route: local}
on_cloud_down: local      # never fail over to cloud
provider_terms:
  max_retention_days: RETENTION_DAYS  # set from your signed terms
  training_use: false
```

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. ### Your demo works and production is stuck.

   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. ### You need senior AI judgement without a full-time hire.

   Fractional AI Lead. Monthly: 90-day minimum, then 30 days' notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. ### An AI feature has to work, by a date.

   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. ### A regulator, an auditor or a customer's security team is asking.

   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL

indica-tech.com/#contact

EMAIL

hello@indica-tech.com

WEBSITE

indica-tech.com

ALL CHECKLISTS

indica-tech.com/checklists/

YOUTUBE

youtube.com/@demotoprod

LINKEDIN

linkedin.com/in/startup-founderx

INSTAGRAM

instagram.com/demotoprod

TIKTOK

tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#local-vs-cloud