@demotoprod · CHECKLIST

Kubernetes probe checklist for AI services

A liveness probe that checks a dependency turns a slow database into a restart storm: every replica fails the same probe in the same window, and the kubelet restarts them all together. The demo never shows it, because the demo ran one pod. Every default and behaviour below is from the Kubernetes documentation; nothing here is a vendor claim. Five checks; the first two are an afternoon's work.

1 Liveness checks only the process

The kubelet restarts a container when its liveness probe fails three times in a row, probing every ten seconds with a one-second timeout by default. Point that probe at an endpoint that calls your database, vector store or model server, and one thirty-second stall fails it on every replica at once, so every replica restarts together. Capacity goes to zero while the dependency was only slow.

- Open every livenessProbe. If its handler touches anything over the network, move that check out.
- Liveness answers one question: can this process respond at all?

2 Dependency checks live in readiness

A failing readiness probe marks the pod unready and takes it out of Service traffic; it does not restart the container. For an AI service that holds a model in memory, that is the difference between a pause and a cold start.

- Give the pod two endpoints: /livez (process only) and /readyz (database, vector store, model loaded).
- The pod leaves traffic while a dependency is down and returns the moment it is back, model still loaded.

```
livenessProbe:  { httpGet: { path: /livez,  port: 8080 } }   # process only, no I/O
readinessProbe: { httpGet: { path: /readyz, port: 8080 } }   # database, vector store,
model loaded
```

3 Slow-loading models get a startup probe

Liveness and readiness checks do not start until the startup probe has succeeded. The documentation's rule: set failureThreshold x periodSeconds long enough to cover the worst-case startup time (its example allows 30 x 10 = 300 seconds).

- Time your worst cold start, weights downloaded, loaded and warmed, and size the startup probe to cover it.
- Do not stretch the liveness probe's initial delay instead.

4 Probe handlers are fast and in-process

An HTTP probe that answers after its timeout counts as a failure, whatever it would have returned. The default timeout is one second.

- Keep the liveness handler constant-time and free of I/O.
- If it cannot answer inside the timeout under peak load, fix the handler before you raise the timeout.

5 A restart storm pages someone

One restart is noise; every replica restarting inside the same minute usually means something they all share, and the events say whether it was the liveness probe. kube-state-metrics exposes restarts per container as kube_pod_container_status_restarts_total.

- Alert when restarts rise on several replicas of one Deployment within the same window.
- Link the alert to the two commands on this page.

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.

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, 30-day 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/#kubernetes-probes