@demotoprod · CHECKLIST

AI failover test

One reliability report counted 2,730 outages across 34 AI and LLM providers in the first half of 2026, about fifteen a day, and a model suspension that ran 18 days 19 hours, taking the platforms built on top of it down with it. A second vendor is not a second failure domain. These three checks take an afternoon; the fourth is a recurring five minutes. Nothing here needs a migration.

1 Find out what your backup really is

Resellers, gateways and clouds collapse into the same upstream more often than the contract suggests: the same model family, served from the same region, behind the same gateway. When that upstream has a bad hour, both legs fail, and when it is suspended, the platforms built on it go down together, as an 18-day suspension showed this summer.

- For each provider write down three things: whose model weights, which cloud and region, which gateway or proxy.
- If any two match the primary, that leg is not a fallback. It is a second route to the same failure.

2 Give every provider its own timeout and retry budget

A single client-side budget is spent by whoever is slow. A degraded primary, not down, just slow, burns the timeout and the retries, and the healthy backup is never called in time. Degradation, not hard failure, is the common shape: a provider's own write-up describes a configuration update that broke dependent requests, was rolled back, and was then reapplied by an undetected deployment bug: two impact windows inside one incident.

- Per provider: its own timeout, its own retry count, its own circuit breaker.
- Cap total attempts across the ladder.
- Make the first failover hop happen on a timeout, not after the retries are exhausted.

3 Test it in production config, five minutes a week

Staging proves the code path, not the dependency graph: staging has different keys, different quotas, different regions and no real traffic. Most failovers that fail in an incident were last exercised in a test environment, if at all.

- Put a five-minute window in the calendar.
- Flip the primary off in production configuration, watch error rate, p95 and cost, then flip it back.
- Write down what the backup's latency and spend were. That number is your failover budget.

4 Decide what "degraded" means before you need it

When both providers are impaired, the only options left are the ones you built earlier: a cached answer, a smaller local model, or a deterministic response that is honest about being one.

- Name the degraded mode per feature, and the signal that turns it on.
- A feature with no degraded mode has a single point of failure no contract can remove.

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/#ai-failover