@demotoprod · SHIP · also LEVERS

Deployment checklist

A notebook runs one cell, once, for one person. With no concurrency, timeout or log, the first burst of traffic or one bad input is an outage with nothing to explain it. Five checks, each sized for an afternoon, cover the service, the container, the logs and the input.

1 Time out every model call in an async handler

Per the HTTP client library's docs, requests do not time out unless a timeout is set explicitly, so code may hang for minutes or more. A blocking call in a coroutine blocks the event loop.

- Run blocking SDK calls with asyncio.to_thread, so one slow call cannot hold the loop.
- Set client connect and read timeouts, and wrap each call in asyncio.timeout() (Python 3.11+).

2 Cap in-flight calls; answer overflow with a 503

Queued requests consume memory and add latency: in the reliability book's worked example, a queue 10 times the thread count at 100 ms per request makes a full-queue request take 1.1 seconds. It advises HTTP 503 past a set number in flight.

- Cap concurrent model calls with a semaphore; load-test at 10, then 50, simultaneous requests.
- Return 503 at once at the cap; if you queue, bound the wait and drop older requests.

```
LIMIT = asyncio.Semaphore(8)       # placeholder
async def ask(prompt):
    if LIMIT.locked():
        raise Overloaded()         # map to HTTP 503
    async with LIMIT, asyncio.timeout(30):
        return await call_model(prompt)
```

3 Ship one pinned container image everywhere

A notebook uses whatever one laptop has installed. The container reference says pulling by digest pins an image to one version, always the same; a pull by tag can fetch a newer one. The installer's hash mode needs every requirement pinned and hashed.

- Build one image from a hash-locked requirements file; run it unchanged everywhere.
- Pin the base by digest; rebuild on a schedule, since a pinned image pulls no security updates.

```
FROM python:<version>-slim@sha256:<digest>
COPY requirements.lock .
RUN pip install --require-hashes -r requirements.lock
```

4 Log every request with its tokens and latency

The twelve-factor method writes each process's event stream, unbuffered, to stdout, so log systems can find past events, graph requests per minute and alert. A GenAI metrics standard, status Development, defines operation-duration histograms and token counters.

- Write one JSON line per request to stdout: id, model, tokens in and out, latency, status.
- Emit gen_ai.client.operation.duration and token-usage counters; alert on error rate.

5 Reject oversized input before it reaches the model

The API security list says to define and enforce a maximum size on incoming parameters and payloads, such as string length and array elements. The LLM risk list repeats it for model inputs and adds timeouts.

- Limit characters per field, items per list and bytes per body; return a 4xx before the model.
- Log each rejection with its reason, so a bad input leaves a trace.

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 SETTINGS BLOCK FOR THE FIVE CHECKS

```
# Names in capitals are yours to set from your own load test
call: {connect_timeout_s: CONNECT_S, read_timeout_s: READ_S}
max_in_flight: MAX_IN_FLIGHT    # past this, answer HTTP 503
image: python:<version>-slim@sha256:<digest>
install: pip install --require-hashes -r requirements.lock
log_to_stdout: [id, model, tokens_in, tokens_out, latency_ms, status]
input_limits:
  max_body_bytes: MAX_BODY_BYTES
  max_field_chars: MAX_FIELD_CHARS
  max_list_items: MAX_LIST_ITEMS
```

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/#notebook-to-production