@demotoprod · KEYS · also LEVERS

Guardrails checklist

In the 2025 top-ten list for LLM applications, prompt injection is risk number 1. A guardrail that filters only the prompt leaves retrieved text and the answer unchecked. Five checks, each sized for an afternoon, cover both sides of the model, its speed and its failures.

1 Put input rails on the prompt and on retrieved text

A 2023 paper demonstrated, on real-world systems, that adversaries can inject prompts into data likely to be retrieved. One open-source toolkit's README says input rails can block a request before it reaches the model, and its PII check can detect or mask sensitive data in inputs, outputs and retrieved content.

- Name the prompt checks (PII, off-topic, jailbreak phrasing); each rejects or masks, and logs a hit.
- Run the same checks on each retrieved chunk and tool result, and mark that text untrusted.

2 Put output rails on the answer: grounding, format, PII

The top-ten list advises scoring responses with the RAG Triad (context relevance, groundedness, question and answer relevance) and validating output formats in code. Its improper-output-handling entry advises treating the model like any other user and encoding its output.

- Reject an answer that fails groundedness against the retrieved text or format against a schema.
- Scan for PII and secrets; encode or parameterize the rest for HTML, Markdown or SQL.

3 Run input rails beside the model

One rails FAQ says prompt checking can run in parallel with generation, the draft discarded if a rail blocks, and gives no deployment-independent latency figure. An agent SDK warns that in parallel mode tokens and tools may already be spent when the tripwire fires.

- Start the check and the model call together; on a block, cancel the call and drop the draft.
- Run the rail first where a started call spends money or acts on records. Time p95 on and off.

```
async def answer(prompt):
    check = asyncio.create_task(input_rail(prompt))
    draft = asyncio.create_task(call_model(prompt))
    if not await check:
        draft.cancel()              # blocked: drop the draft
        return SAFE_REPLY
    return await draft
```

4 Fail closed on high-risk paths

The same FAQ says no library-wide failure policy applies to every rail, and a safe response should be returned when a control cannot decide. A weakness entry traces failing open to wanting to fail functional. Streamed text can reach the client before a late block.

- For each high-risk path, define a rail error or timeout as a block with a fixed safe reply.
- Hold streamed output until the output rail has read each chunk; force a rail timeout in staging.

5 Red-team the rails every week

The top-ten list advises regular penetration testing and breach simulations that treat the model as an untrusted user. A paper drawing on over 100 generative AI products says the work of securing AI systems will never be complete.

- Keep a versioned set of attack prompts, poisoned documents and legitimate prompts; run it weekly and after any change.
- Log bypass rate and false-block rate per run; fix until both meet limits your team has written down.

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

A RAILS POLICY TO ADAPT AND PASTE

```
input_rails: [pii, off_topic, jailbreak]       # on the prompt
retrieval_rails: [pii, mark_untrusted]          # each chunk, tool result
output_rails: [groundedness, schema, pii, encode]
run_input_rails: parallel     # rail first where a call spends or acts
on_rail_error: block          # high-risk paths fail closed
safe_reply: SAFE_REPLY
hold_stream_until_checked: true
red_team:
  schedule: weekly            # and after any change
  max_bypass_rate: <your limit>
  max_false_block_rate: <your limit>
```

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/#guardrails