@demotoprod · BUDGET · also LEVERS

Agent cost ceiling checklist

In one agent SDK's documentation the turn limit and the spend limit both default to no limit. A threat-intelligence firm's September 2026 special report gives its own account of one case: an agent looping on a corrupted value, over 15,000 reasoning API calls in under an hour and a billing spike of about $50,000. Five checks put a budget, a circuit breaker and a step limit in your own code, then back them up at the provider.

1 Enforce the daily budget in your own code, before each call

Provider spend limits are monthly and enforcement is not instantaneous. An SDK budget cap compares estimated spend after each response, so spend can pass it by up to one response plus running subagents. A daily ceiling is a counter you check before every call.

- Keep one counter per task per day in a store that survives a restart; before each call add its worst-case cost and refuse if the total would pass the cap.
- When the guard refuses, stop and file a ticket; the agent can never raise its own cap.

2 Trip a circuit breaker on consecutive failures, and reset it by hand

A circuit breaker blocks a resource after failures reach a threshold instead of retrying. The documented counter is time based, so it trips only when enough failures land inside one interval. The special report asks for a halt after a set number of consecutive failures and gives no number.

- Count failures per task in a row with no time window, count a rejected output as a failure, and reset only after a success.
- When it trips, stop the task, log the last failing input, notify a person, and keep it open until a person closes it.

3 Cap the steps per task, and handle the cap as a logged ticket

In one agent SDK the turn limit also counts the final reply and defaults to no limit. In a widely used orchestration library the tool-call limit by default blocks excess calls and lets the model go on; its model-call limit ends the run. Test what yours does when hit.

- Cap tool calls per task and sub-task depth from your longest legitimate run in the logs.
- Handle the cap as a logged ticket (task id, step count, last inputs) outside any retry wrapper, and force a loop in staging to confirm one ticket.

4 Set a hard spend limit on each agent's project or workspace

One provider separates a spend alert, which notifies while traffic continues, from a hard limit, where requests return a 429. Another provider's monthly cap returns an error typed like a rate limit, with no retry-after header, and retries fail until access resumes.

- Give each agent its own project or workspace with a hard limit above your in-code cap, so your code trips first.
- Mark the limit errors terminal in the retry policy (project_spend_limit_exceeded; enforced_spend_limit_reached, or a 400 for a limit you set yourself), and rehearse with a small limit on staging.

5 Alert at 50 and 90 percent of each cap, from your own counter

In the reported incident the agent made over 15,000 calls in under an hour. One cloud provider's guide says billing lags usage, that an alerts-only budget does not cap spending, and its console proposes alerts at 50%, 90% and 100%.

- Alert at 50 and 90 percent of each task's cap from your own counter, so it does not wait for billing; name who acts and what they do first.
- Set the cloud account's budget alerts at 50, 90 and 100 percent as a second signal.

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 COST POLICY TO ADAPT PER AGENT

```
# Names in capitals are yours to set from your own logs
budget:
  per_task_per_day_usd: DAILY_CAP_USD
  check: before_each_call     # refuse if worst-case cost passes the cap
  alert_at_percent: [50, 90]  # from your own counter
breaker:
  trip_after_failures_in_a_row: FAILURE_THRESHOLD
  reset: by_hand
steps:
  max_tool_calls_per_task: MAX_TOOL_CALLS
  on_cap: open_ticket         # outside any retry wrapper
terminal_errors: [project_spend_limit_exceeded, enforced_spend_limit_reached]
```

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/#agent-cost-ceiling