Maxence Foulon

Free resource · Checklist

AI Plan B: 8 safeguards so an AI project survives an outage, a price hike or a market crash

In 15 minutes, you'll know what stops at your company if your AI provider cuts you off tomorrow. In one day, you'll have a plan so it doesn't stop you.

Reading time 10 min 16 verified sources Updated November 7, 2026

The 30-second version

- **Anthropic lost $104 billion over nine months, mostly non-cash accounting charges, and some investors doubt its IPO will happen in November. On November 5, the Nasdaq had its worst day since April 2025.[1, 2, 3]**
- **The concrete risk for a customer is dependency: in June 2026, Claude Fable 5 was switched off for every customer for almost three weeks after a US government decision.[4, 5, 6]**
- **The fix comes down to 8 safeguards. The first one takes 15 minutes: the inventory. For a backup, Mistral released the weights of Mistral Large 4 on October 29: far behind the leaders, but downloadable.[7, 8]**

Context

What's behind the word “bubble”

In October 2026, three signals collided. CoreWeave, the best-known of the GPU rental companies dubbed “neoclouds”, saw its bonds trade at yields close to 15%, and its stock lost half its value between October 1 and November 5.[9, 10] Elon Musk predicted that most of the 200-plus

neoclouds would die or be bought out because credit had dried up.[11] And Anthropic filed for its IPO, reporting an $80.7 billion net loss over six months.[4]

That last number is scary, but easy to misread. Of that, $72.4 billion comes from revaluing financial instruments, mostly notes that convert into shares: the more the company is worth, the more they weigh on the books, without a single dollar leaving the bank. The operating loss was $693 million, on $16.3 billion of revenue.[4]

In early November, the pressure went up a notch. Anthropic's updated IPO filing (the S-1/A, filed with the SEC) adds a $23.3 billion net loss for the third quarter, again mostly non-cash ($21.9 billion of revaluation), on $17 billion of revenue.[1] On November 5, two days after the US midterm elections, the Nasdaq had its worst day since April 2025 (–5.5%), and potential investors told Reuters they doubt Anthropic's IPO will happen in November.[2, 3]

The real weight

Compute commitments. Anthropic has signed roughly $895 billion of purchase and lease agreements, almost all of it compute and infrastructure: $447.5 billion before the end of 2030, the rest through 2045.[1] That bet, and the bet of the whole chain financing GPUs, is what's really at stake.

Why this is your problem

You don't pay CoreWeave or Nvidia. But your AI provider rents compute from that chain, and if it's a US company, it answers to US law. In June 2026, the Commerce Department imposed export restrictions on Fable 5 and Mythos 5. Unable to sort its customers by nationality, Anthropic switched both models off for everyone on June 12, two days after Fable 5 launched. Access was restored on July 1.[4, 5, 6] Earlier that year, on February 27, the US President had ordered the federal government to stop using Anthropic's models.[1] Its IPO filing (the S-1) warns that such export restrictions could recur, last longer or extend to other models.[4]

01

Take inventory (15 minutes)

Before choosing anything, find out where you depend on AI. List every use of AI on your team, including the ones nobody has reported (shadow AI): the finance team's browser extension, the sales rep's assistant, an intern's script.

- [ ] For each use: who uses it, for which task, with which tool and which model.
- [ ] What happens if it disappears for a week: nothing, friction or a full stop.
- [ ] Where the data goes: which provider, which country. If you don't know, write “ask”.

AI dependency sheet (paste into a spreadsheet)

```
Use | Team | Tool | Model | Provider | Data location | If down 1
week | Backup tested? | Owner
Example: support replies | Customer support | Ticketing tool + API |
(model) | (provider) | ask | Full stop | No | (first name)
```

02

Test a backup model, don't just name one

During the Fable 5 shutdown, Anthropic's other models, including Opus 4.8, stayed available.[5] A team that had already run its tasks on another model could switch with a config change. A team that hadn't done so discovered the differences in production. Fable 5 had only just launched, but the mechanism applies to any model.

- [ ] Pick two backups: another model from the same provider (for a shutdown that targets one model) and a model from another provider (for a wider outage, a price hike or a government decision).
- [ ] Run your test set (see 04) on both and write down the gaps.
- [ ] For sensitive data, keep an option hosted in a jurisdiction you trust (in Europe if GDPR applies to you) or on your own servers.

03

A single entry point in your code

If the model name is hard-coded in forty files, switching providers takes weeks. If every call goes through one internal function and the model is set in configuration, the technical switch takes an hour; then you check the results with your test set (see 04).

- [ ] Every AI call goes through one internal module.
- [ ] The main model and the backups live in configuration, not in code.
- [ ] Prompts are stored separately, versioned, with the model they were tested on.
- [ ] A toggle (feature flag) lets you fail over without redeploying.
- [ ] The backup provider is already approved for your data (contract, country, security) before any automatic failover sends it anything.

Configuration template (adapt it)

```
ai:
  primary:
    provider: "to fill in"
    model: "to fill in"
  backup_same_provider:
    model: "to fill in"
  backup_other_provider:
    provider: "to fill in"
    model: "to fill in"
  failover:
    automatic_if: "more than 5% errors over 10 minutes"
    failover_owner: "owner's name"
  budget:
    monthly_cap_usd: "to fill in"
    alert_at: "80%"
```

04

Your own test set

Vendors publish the benchmarks they shine on. Moonshot, Kimi's Chinese developer, shows Kimi K3.1 Pro 0.7 points behind Claude Opus 5.5 on Terminal-Bench 4.0 (65.7 vs 66.4), at half the price: $2 vs $4 per million input tokens, $10 vs $20 for output.[12, 13, 14] That tells you nothing about your contracts, your tickets or your brand voice. Twenty real cases will.

- [ ] Collect 20 real, anonymized requests with the answer you expect.

- [ ] Add 5 trap cases: missing information, out-of-scope request, personal data that must not be repeated.
- [ ] Score every candidate model with the same rubric, every time a model or a price changes.

Prompt: turn your real cases into a test set

```
Here are 20 real requests from my team, anonymized, with the answer
we consider good.
For each one:
1. List what the answer must contain (3 checkable criteria at most).
2. State what would make it unacceptable (error, tone, made-up
information).
3. Suggest a 0 to 2 score per criterion.
Then add 5 trap cases based on these requests: missing information,
out-of-scope request, personal data that must not be repeated.
Return everything as a table: Case | Input | Criteria | Unacceptable
if | Scoring.

[Paste your 20 requests here]
```

05

The questions to ask before you sign

Your contract is your only legal protection against a price hike or a model being retired. Anthropic's own IPO filing (S-1) mentions possible pricing changes, such as outcome-based or per-task pricing.[4] I'm not a lawyer: have a lawyer review the clauses. Here are the questions to send your provider.

Email to send your AI provider

```
Hello,

Before we sign or renew our contract, could you confirm in writing:
1. The notice period for a price increase, and whether there is a
cap.
2. The notice period before a model we use is retired, and how long
the migration window is.
3. How we export our data, prompts and logs: format and timeline.
4. Which countries our data is processed and stored in, and by which
subprocessors.
5. What happens to our service if an export restriction or a
government decision forces you to switch a model off: notice,
replacement model, refund.
6. Your uptime commitments (SLA) and the service credits that apply.

Thank you,
(first name)
```

06

A budget that can absorb a price hike

Prices move. In September 2026, Anthropic cut the cost of typical workloads by about 25% with Fable 5.1, and the cost per task by about 40% with Opus 5.5.[4] But those cuts happen in a market where providers commit to hundreds of billions in compute and the companies renting them GPUs borrow at high rates.[1, 9] Nothing guarantees they'll last.

- [ ] Measure cost per task (per ticket handled, per document produced), not just the monthly bill.
- [ ] Work out what happens to your project if the price doubles. If it breaks, find a cheaper model that's good enough.
- [ ] Set a spending cap and an alert with every provider.

07

An open option, after reading the license

An open-weight model that you host yourself or with a host you choose doesn't depend on its vendor's decisions once downloaded. But “open” doesn't mean “no strings attached”. The Kimi K3.1 Pro license, for example, requires a separate agreement if you resell it as a service with more than $20 million in revenue over 12 months; internal use is exempt.[15]

On the European side, Mistral released the weights of Mistral Large 4 on October 29, under a modified MIT license that allows commercial use; only running a business that resells access to the model requires a separate license.[7, 16] It is far behind the leaders: a score of 26 on the Artificial Analysis Intelligence Index, an independent ranking, on November 7, versus 46 for Kimi K3.1 Pro and 57 to 60 for the best US models.[8] At about a trillion parameters, in practice it needs a hosting provider; it won't run on a small company's server.[16] A backup doesn't need to be the best: it needs to handle your critical tasks, and only your test set will tell you.

- [ ] Find an open model that can handle your critical tasks, even if less well.
- [ ] Read its license: commercial use, resale, required attribution.
- [ ] Run it through your test set and note how far it falls behind your main model.

08

A written outage plan, tested every quarter

The day your model disappears, nobody should be asking who decides. One page is enough.

AI outage plan (one page)

```
Trigger: the main model is unavailable, restricted, or its price
rises by more than ___%.
Decision owner: ___ (backup: ___).
Time to decide: ___ hours.
Failover: switch to ___ (same provider), then to ___ (other
provider) if needed.
What stops during the switch: ___.
Message to teams: “The ___ assistant is running on a different model
until ___. Report any odd answer to ___.”
Test: once a quarter, run one day (or part of the traffic) on the
backup and compare results.
```

- [ ] Write the page and share it with the teams involved.
- [ ] Schedule the first failover test within a month.

Sources

1. SEC EDGAR, Anthropic, PBC, Form S-1/A https://www.sec.gov/Archives/edgar/data/2130357/000162828026068676/anthropicpbc-sx1ax1.htm, November 2, 2026.
2. Reuters, via Investing.com, S&P 500 futures steady after midterm rally; Nasdaq lags on chip weakness https://www.investing.com/news/stock-market-news/sp-500-futures-steady-after-midterm-rally-nasdaq-lags-on-chip-weakness-5003237, November 5, 2026.
3. Reuters, via Investing.com, Exclusive: AI backlash could threaten Anthropic IPO timing, potential investors say https://www.investing.com/news/stock-market-news/exclusiveai-backlash-could-threaten-anthropic-ipo-timing-potential-investors-say-5006110, November 5, 2026.
4. SEC EDGAR, Anthropic, PBC, Form S-1 https://www.sec.gov/Archives/edgar/data/2130357/000162828026067168/anthropicpbc-sx1.htm, October 26, 2026.
5. Greenberg Traurig, AI Company Anthropic Suspends Access to Claude Fable 5, Claude Mythos 5 Following US Export Control Directive https://www.gtlaw.com/it/insights/2026/6/ai-company-anthropic-suspends-access-to-claude-fable-5-claude-mythos-5-following-us-export-control-directive, June 17, 2026.

6. Engadget, US government allows Anthropic to redeploy its Mythos and Fable AI models https://www.engadget.com/2205599/anthropic-redeploy-mythos-fable-ai-models/, June 30, 2026.
7. Mistral AI, Mistral Large 4 (Le Chonk) is now open weight https://mistral.ai/news/mistral-large-4-oss/, October 29, 2026.
8. Artificial Analysis, LLM Leaderboard, Intelligence Index v5.0.1 https://artificialanalysis.ai/leaderboards/models, accessed November 7, 2026.
9. CNBC, CoreWeave shares tumble as bond yields near 15% on concern about AI borrowing binge https://www.cnbc.com/2026/10/14/coreweave-stock-bond-yields-ai-debt.html, October 14, 2026.
10. StockAnalysis, CoreWeave (CRWV) stock price history https://stockanalysis.com/stocks/crwv/history/, accessed November 6, 2026.
11. CNBC, Elon Musk says more than 200 neoclouds face 'guppy-level death' https://www.cnbc.com/2026/10/26/neocloud-stocks-today-coreweave-nebius-iren-elon-musk.html, October 26, 2026.
12. Moonshot AI, Kimi K3.1 Pro https://www.kimi.ai/blog/kimi-k3-1-pro, October 23, 2026.
13. Artificial Analysis, Kimi K3.1 Pro https://artificialanalysis.ai/models/kimi-k3-1-pro, accessed November 6, 2026.
14. Anthropic, Pricing (Claude API) https://platform.claude.com/docs/en/about-claude/pricing, accessed November 6, 2026.
15. Moonshot AI, Kimi K3.1-Pro License https://huggingface.co/moonshotai/Kimi-K3.1-Pro/raw/main/LICENSE, Hugging Face, accessed November 6, 2026.
16. Hugging Face, Mistral AI, Mistral-Large-4-1T-A52B https://huggingface.co/mistralai/Mistral-Large-4-1T-A52B, accessed November 7, 2026.

Want an outside look at your dependency sheet?

I take freelance projects: I design and ship AI tools wired into the software a team already uses. Send me your sheet and we'll look at where it would break.

Tell me about your project **See my work**

KakaoTalk, WhatsApp or an email to copy: maxence.pro/#contact

More free resources: maxence.pro/resources/

© 2026 Maxence Foulon · maxence.pro · Free to share with credit.