THE BUILDER'S STACK · THE PRACTICE · SECURITY

The Security Posture

The case that your AI product ships defended: the threats you named, the keys it holds, the copies your data makes, the chain you imported, the layers that stop what the prompt cannot, and the red-team findings now pinned as eval cases. Fill it from work that exists and runs, sign it, and re-open it whenever the product changes underneath.

PRODUCT / FEATURE

OWNED BY

MODEL AND STACK, ONE LINE

DATE

1 The job and the doors

The job in one sentence, then every place outside text gets in. You cannot defend doors you have not listed.

A USER HIRES THIS PRODUCT TO...

THE DOORS: TEXT BOX, FILES, RETRIEVED PAGES, TOOL RESULTS, VENDOR TOKENS

2 The threat map

Five lines, one each: an asset, a door, and the path between them. Plain sentences a teammate can argue with.

TOP FIVE THREAT LINES

THE ONE YOU FEAR MOST, AND WHY

3 The authority table

Every action, the identity it borrows, the scope it carries, its rung on the ladder. The keys, not the model, set the blast radius.

ACTION -> BORROWED IDENTITY -> SCOPE -> RUNG

- [ ] Every key can be revoked in minutes, and we know the steps

- [ ] One revocation rehearsed, with the date written

4 Data and retention

Typed, stored, embedded, logged. Every stop is a copy; every copy gets a retention number and a reader list.

TYPED: RETENTION (DAYS)

STORED: RETENTION (DAYS)

EMBEDDED: RETENTION (DAYS)

LOGGED: RETENTION (DAYS)

- [ ] The tenant wall was tested as another customer
- [ ] Logs are scrubbed of secrets and personal data

5 The supply-chain register

Every part you imported: weights, hosting, libraries, tools, MCP servers. Publisher, version, pin.

PART -> PUBLISHER -> VERSION -> PINNED?

- [ ] Weights load from named publishers in safe formats
- [ ] Every dependency is pinned; nothing updates itself into production

6 Defense in layers

For each top threat: the deterministic stop, the classifier, the human gate. The prompt is a request; the floor is a fact.

THREAT -> DETERMINISTIC STOP -> CLASSIFIER -> HUMAN GATE

- [ ] No threat on the map is defended by prompt wording alone
- [ ] Spend caps are live: per user and per day, with an alert and a hard stop

7 The red-team ledger

Attack, finding, eval case, gate. A finding is finished when it is a case the regression gate runs.

FINDING -> EVAL CASE ID -> GATE STATUS

ATTACK-HOUR CADENCE

LAST SESSION DATE

- [ ] Every red-team finding above is an eval case in the regression gate

8 The kill switch and the first 24 hours

Shut the door, revoke the keys, read the receipts, tell the people. Rehearsed once on purpose.

KILL SWITCH: WHO PULLS IT

TIME TO OFF (REHEARSED)

THE FIRST 24 HOURS, IN STEPS

- [ ] The kill switch was pulled once on purpose, and the time is written above

9 Chained artifacts

The Posture is filled from artifacts that exist and run, never from intentions.

- [ ] The Agent Charter is signed; the authority table above extends it
- [ ] The Quality Bar is filled and its suite, red-team cases included, runs in the gate

10 Sign-offs and review

Named people, dated lines. Decide now what re-opens this page.

BUILT BY

CHALLENGED BY

APPROVED BY

LAUNCH DATE

WHAT RE-OPENS THIS PAGE (NEW TOOL, NEW CORPUS, NEW MODEL, NEW KEY)