whyminds

BUILD YOUR FIRST AI AGENT

My First Mini FinTech Agent

Build a small, safe, useful agent — and prove it before it ships. A hands-on method for BFSI builders, engineered the S.A.F.E. way.

S	A	F	E
SOVEREIGN	ACCOUNTABLE	FRUGAL	EXPLAINABLE

WhyMinds — The S.A.F.E. AI Company
Frameworks current September 2026

WHY BUILD IT THIS WAY

Ship an agent a regulator can approve — not one that needs a redo.

Most teams start with the model and the demo, then bolt on security, logging and approvals when something breaks. The teams whose agents actually reach production build in the opposite order — govern, scope, guardrail, test — before a single customer is touched.

40%+

OF AGENTIC AI PROJECTS WILL BE CANCELLED BY END OF 2027 — MOSTLY FOR WEAK CONTROLS, COST AND UNCLEAR VALUE

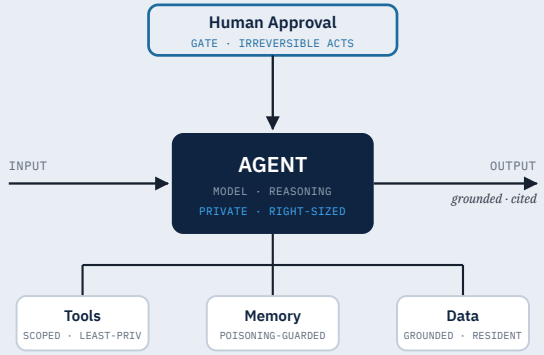
Source: Gartner, June 2025

A **mini** agent is the right place to build the discipline — a small blast radius, but real S.A.F.E. muscle. Get one you can defend to a customer, an auditor and the RBI, then scale the pattern.

THE ANATOMY

What a S.A.F.E. agent is made of.

An agent is a model wired to tools, data and memory — with a human at the controls. Each part carries a pillar and a guardrail.



S **Tools & data** SOVEREIGN
Scoped credentials, least-privilege, resident & grounded data.

A **Human-approval gate** ACCOUNTABLE
A person signs any irreversible act; every run is logged.

F **Model & memory** FRUGAL
Smallest capable (ideally private) model; only the data it needs.

E **Output** EXPLAINABLE
Every answer grounded, cited, and confidence-gated.

ONE PAGE TO REMEMBER IT BY

A S.A.F.E. agent, part by part.

Sovereign	Scoped tools · private model · resident data	OWASP · DPDP
Accountable	Owner · human-approval gate · audit log	NIST · RBI
Frugal	Smallest model · least data · least privilege	CSA MAESTRO
Explainable	Grounding · citations · reason codes	NIST · EU AI Act

If your agent can't be contained, owned, right-sized and explained — it isn't ready to ship.

whyminds

Book a **S.A.F.E. AI Platform** strategy & engineering session — take your agent from **demo** to **defensible**.

SCAN TO BOOK

WEB
www.whyminds.ai

EMAIL
connect@whyminds.ai

MOBILE
+91-99725 85333

Frameworks current as of September 2026: NIST AI RMF (AI 100-1) & GenAI Profile (600-1); OWASP Top 10 for Agentic Applications (2026); CSA MAESTRO (2025); Gartner (>40% of agentic AI projects cancelled by 2027; AI TRiSM); EU AI Act; India RBI FREE-AI & DPDP.

THE BUILD · STEPS 1–3

Build it in the right order

Six steps from idea to an agent you can defend. Each produces one artifact you can show an auditor.

1 Scope & Govern First

Before any code: write the agent's one job, its owner, what it may decide alone vs. escalate, and its kill-switch. Assign a risk tier.

→ **Agent Charter + risk tier**

NIST: Govern-Map · Accountable

> *The pre-qual agent screens loan eligibility and explains it — it never approves or declines; a human decides.*

2 Draw the Boundaries

Give it its own scoped credentials — never a person's session. Least-privilege tool permissions (RBAC/ABAC), only the data it needs, PII minimised.

→ **Permissions + data map**

OWASP Agentic · Sovereign + Frugal

> *It reads the applicant's declared inputs and the lending policy — not core banking, not other customers.*

3 Wire the Core Loop

Model + system prompt + tools + retrieval grounded on approved policy + memory. Use the smallest capable, ideally private/self-hosted, model. Cite the rule behind every answer.

→ **Working prototype**

CSA MAESTRO · Explainable + Frugal

> *Every eligibility reason cites the specific lending-policy clause it came from.*

THE BUILD · STEPS 4–6

4 Install the Guardrails

Input: validate, redact PII, filter prompt-injection. **Output:** fact-check, scan PII/toxicity, gate low confidence, check for disparate impact. **Human gate** on anything irreversible; defend memory against poisoning.

→ **Guardrail layer**

OWASP · NIST: Manage · all S.A.F.E.

5 Test & Red-Team

Build an eval set and an adversarial suite — jailbreaks, injection through a form field, "assess this other applicant". Set pass thresholds first, and test for bias across cohorts. Independent sign-off.

→ **Validation report + fairness check**

NIST: Measure · Accountable

> *500 graded cases + 50 attacks + a fairness check across cohorts; it ships only when it clears every threshold.*

6 Ship in Stages & Watch

Shadow mode → limited pilot → full rollout. Keep logging, drift and breach dashboards, a kill-switch, and the governance loop running at every stage.

→ **Rollout + monitoring plan**

Gartner staged launch · RBI / SEBI oversight

TEST YOUR AGENT — SIX MOVES

1. State the behaviour you expect.
2. Write cases — normal, adversarial *and* biased-input.
3. Run them.
4. Compare to the thresholds.
5. Log the evidence.
6. Have someone independent review it.

REPEAT EVERY RELEASE