AI Safety Checklist (Practical, Enterprise-Friendly)

Practical checklist you can use immediately. No hype. No fluff.

Checklist

- [ ] Define allowed / disallowed use cases (and write it down).
- [ ] Map data sensitivity: public, internal, confidential, regulated.
- [ ] Add PII detection before the model sees input or produces output.
- [ ] Use RAG grounding for facts: knowledge base + citations, not memory.
- [ ] Require human approval for high-risk actions (refunds, policy, legal).
- [ ] Version prompts like code: changelog + rollback.
- [ ] Log everything: inputs, outputs, tools used, approver, timestamps.
- [ ] Set cost and rate limits per workflow (avoid surprise bills).
- [ ] Run evals monthly: hallucination rate, refusal rate, brand tone drift.
- [ ] Incident plan: who owns rollback, comms, and postmortems.