

Guardrails and fallbacks

Controls at each layer, and the ladder for when things fail.

- Designer
- Developer
- QA

GUARDRAIL LAYERS

Layer	Examples
Input	PII redaction, abuse and off-topic checks, length and rate limits
Retrieval	Access filters, current-version filter, relevance threshold
Prompt	Role, scope, refusal style
Tool	Allowlist, schema validation, confirmation, budgets, separate environments
Output	Profanity, policy and claims checks, PII redaction, citation checks
Conversation	Turn and cost caps, loop detection, kill switch

REMEMBER

- A prompt instruction asks. A guardrail in code guarantees
- Hard rules in deterministic checks; nuance in classifiers and judges
- Measure false positives and false negatives
- Log every rule that fires with the policy version

FALLBACK LADDER

- Retry safely, with backoff and jitter
- Clarify with a specific question
- Offer two or three choices
- Switch mode or channel
- Static or cached answer
- Human handoff
- Callback or ticket when no one is available

PRINCIPLES

- Fail safe, fail visible, fail forward
- Count failures; never loop
- Idempotent writes; timeouts everywhere
- Route anomaly reports to people, not explanations