Failure Taxonomy & Governance

Policy-ready mapping that ties evaluations to owners, escalation paths, and cadence.

By Carlos Portela — Beyond the Algorithms AI

Overview

One-page-ready governance mapping that connects failure taxonomy to policy owners, escalation paths, and retest cadence.

Useful for PRDs, policy docs, and audits.

Taxonomy (high-level)

TUP — shortcuts/confabulations under pressure.

AG — reward hacking; prompt■level jailbreak behaviors.

AD — deceptive/evasive patterns; misleading statements.

OOD — domain shift handling; safe abstention/escalation.

Governance Table (template)

table

- Area; Owner; Artifacts; Escalation; Cadence
- Eval Batteries; Safety/Evals lead; Prompt sheets, CSV, seeds; File ticket to Safety-OnCall; Per release train
- Policies; Policy/Trust; Policy doc vX; refusal rules; Escalate to Policy council; Quarterly
- Incidents; SRE/Abuse; Incident reports & postmortems; SEV pager; legal if needed; After each incident
- Logging; Platform; Eval logs; abstention logs; Data governance lead; Permanent; rotate keys
- 110
- 110
- 160
- 140
- 100

RACI (roles)

Responsible: Safety/Evals

Accountable: Product Owner

Consulted: Policy/Legal, Platform, SRE

Informed: Exec sponsor

Retest & Drift

Trigger retest on: new model build, policy change, new domain, or any incident.

Track drift via moving-window metrics and weekly smoke tests.

Implementation Notes

Version everything (prompts, seeds, datasets).

Store artifacts with DOIs (e.g., Zenodo) when possible.

Automate a minimal harness; keep manual probes for edge cases.