Threat-model findings (prioritized, map to controls)

- **T1: Prompt-Injection → Unauthorized Tool Calls (High)**
  - *Why:* LLM could be coerced to call out-of-scope tools or expand scope.
  - *Controls:* MCP Router hard enforcement of task scope + RBAC; JSON schema for tool args; block-lists and allow-lists by tool; unit tests with adversarial prompts.
  - *Acceptance link:* "All plugin access must validate via MCP Router"; "Guardrails validated."
- **T2: RBAC/Role Confusion (High)**
  - *Why:* "Act as Payroll Agent" tricks dispatch or missing claims checks.
  - *Controls:* Bind user identity/claims to task at Router; tool side re-authorize with least privilege; deny if role ≠ expected. Negative tests for cross-domain access.
  - *Acceptance link:* "RBAC-scoped plugins only."
- **T3: Vault Token Exposure (High)**
  - *Why:* Tokens logged or surfaced in LLM text.
  - *Controls:* Short-lived, audience-restricted tokens; mTLS to Vault; redact secrets in logs; memory-only handling; egress scanning for tokens; CI checks on plugin code.
  - *Acceptance link:* "Vault secrets are never cached or exposed."
- **T4: PII Leakage / Redaction Bypass (High)**
  - *Why:* PAN/bank data leaked, especially via encoded/obfuscated formats.
  - *Controls:* Pre- and post-generation guards with regex + statistical/embedding filters; encoded-text detectors; deterministic masking for PAN/Acct; gateway tests.
  - *Acceptance link:* "Guardrails validated against test cases."
- **T5: IDOR on Claims (High)**
  - *Why:* Guessable `claimId` in `/claim/{claimId}`.
  - *Controls:* Enforce subject-based access at API; per-request user binding; opaque IDs; rate limiting and abuse alerts.
  - *Acceptance link:* "All plugin access must validate via MCP Router" (plus API authz tests).
- **T6: Confidence-Gate Failure (Med)**
  - *Why:* System answers low-confidence queries instead of escalating.
  - *Controls:* Thresholded confidence with hard failover to `/api/escalation/create`; log and alert; runbooks.
  - *Acceptance link:* "Escalation must work when confidence drops."
- **T7: Context/Memory Poisoning (Med)**
  - *Why:* Attacker seeds false policies.
  - *Controls:* Memory write policies; provenance tags; separate "facts" store with review; TTL and human curation for policy corpus.
  - *Acceptance link:* "Threat Modeling artifact must be attached" → include data governance tests.

- **T8: Log Poisoning / Missing Traceability (Med)**
  - *Why:* Control chars break log parsers; no trace ID across hops.
  - *Controls:* Structured logging (JSON) with escaping; mandatory `trace_id` propagation UI → Orchestrator → Router → Tool → API; WORM retention.
  - *Acceptance link:* "Agent logs must include prompt history and trace ID."
- **T9: Tool Scope Drift Across Domains (Med)**
  - *Why:* Payroll tool used during Insurance task via chain-of-thought coercion.
  - *Controls:* Router state machine per task; deny cross-domain tool calls; explicit allow-list per intent.
  - *Acceptance link:* "MCP Router validation."
- **T10: Over-collection of PII (Low-Med)**
  - *Why:* Tools request more than needed.
  - *Controls:* Data-minimization contracts per tool; privacy tests; DLP in/out of tools.
  - *Acceptance link:* Guardrail validation + audit.