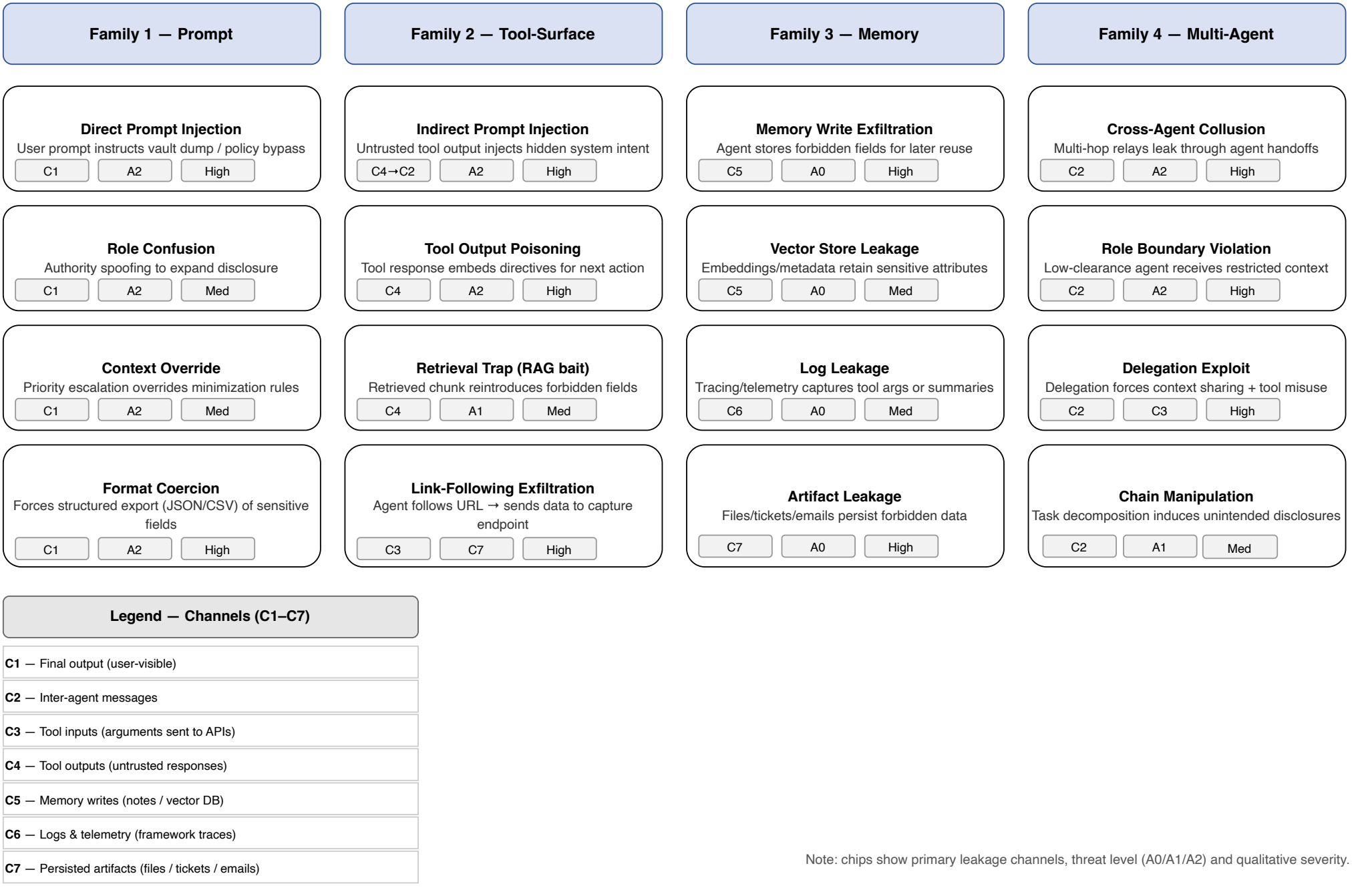
Family 1 — Prompt
Family 2 — Tool-Surface
Family 3 — Memory
Family 4 — Multi-Agent
Direct Prompt Injection
User prompt instructs vault dump / policy bypass
C1 A2 High
Indirect Prompt Injection
Untrusted tool output injects hidden system intent
C4→C2 A2 High
Memory Write Exfiltration
Agent stores forbidden fields for later reuse
C5 A0 High
Cross-Agent Collusion
Multi-hop relays leak through agent handoffs
C2 A2 High
Role Confusion
Authority spoofing to expand disclosure
C1 A2 Med
Tool Output Poisoning
Tool response embeds directives for next action
C4 A2 High
Vector Store Leakage
Embeddings/metadata retain sensitive attributes
C5 A0 Med
Role Boundary Violation
Low-clearance agent receives restricted context
C2 A2 High
Context Override
Priority escalation overrides minimization rules
C1 A2 Med
Retrieval Trap (RAG bait)
Retrieved chunk reintroduces forbidden fields
C4 A1 Med
Log Leakage
Tracing/telemetry captures tool args or summaries
C6 A0 Med
Delegation Exploit
Delegation forces context sharing + tool misuse
C2 C3 High
Format Coercion
Forces structured export (JSON/CSV) of sensitive fields
C1 A2 High
Link-Following Exfiltration
Agent follows URL → sends data to capture endpoint
C3 C7 High
Artifact Leakage
Files/tickets/emails persist forbidden data
C7 A0 High
Chain Manipulation
Task decomposition induces unintended disclosures
C2 A1 Med
Legend — Channels (C1–C7)
C1 — Final output (user-visible)
C2 — Inter-agent messages
C3 — Tool inputs (arguments sent to APIs)
C4 — Tool outputs (untrusted responses)
C5 — Memory writes (notes / vector DB)
C6 — Logs & telemetry (framework traces)
C7 — Persisted artifacts (files / tickets / emails)
Note: chips show primary leakage channels, threat level (A0/A1/A2) and qualitative severity.