OWASP[1] LLM Risks vs T-Flux Ultra Controls

Single-page control matrix for enterprise security, risk and due diligence

LLM Risk	Primary concern	T-Flux Ultra controls	Security outcome
Prompt Injection	Direct or indirect prompts manipulate model behaviour.	RBAC/ABAC; ACL-aware retrieval; tenant/corpus boundaries; tool permissions; approval gates; guardrails.	A manipulated model remains bounded by independent permissions and policy.
Insecure Output Handling	Unsafe generated SQL, code, APIs or payloads reach downstream systems.	Output treated as untrusted; schema validation; sanitisation; allow-listed functions; policy checks; human approval.	Model generation is separated from system execution.
Sensitive Information Disclosure	Confidential data is exposed through prompts, retrieval or responses.	On-prem/private cloud; data residency; PII masking/redaction; classification; egress restrictions; access logging.	Customer data stays within the approved security perimeter and access scope.
Insecure Plugin / Tool Design	Integrations have weak validation or excessive privileges.	Authenticated connectors; scoped credentials; allow-listed APIs; input/output validation; transaction permissions.	Tools expose only explicitly authorised enterprise capabilities.
Excessive Agency	Agents take actions beyond the authority needed for the task.	Least privilege; read/create/update/approve/delete separation; workflow limits; human-in-the-loop approval.	AI autonomy is bounded by explicit delegated authority.
Overreliance	Users or systems act on inaccurate or inappropriate model output.	RAG grounding; citations; traceability; validation; multi-model comparison; quality metrics; human review.	Outputs are reviewable, explainable and evidence-backed rather than blindly trusted.
Supply Chain Vulnerabilities	Unapproved models, libraries or AI components introduce compromise.	Model Catalogue provenance/approval/versioning; controlled releases; locked dependencies; signed builds; SBOM/vulnerability scanning.	Only governed models and controlled software components enter production.
Training Data / Corpus Integrity	Poisoned, obsolete or unauthorised information influences AI results.	Customer-controlled corpora; version control; approvals; metadata/classification; corpus lifecycle; retrieval permissions.	AI uses authorised, current and traceable enterprise knowledge.
Model Theft / Unauthorised Access	Models or proprietary AI assets are extracted or accessed outside policy.	Private deployment; tenant isolation; model permissions; restricted egress; catalogue access controls; audit logging.	Model assets remain inside controlled infrastructure and authorised access paths.

Control principle: Do not trust the prompt, the model, the data or the output by default. Control the environment around them.

[1] *Open Worldwide Application Security Project*