Responsible AI and LLM Security

The trouble with Black-Box Intelligence

Classical Software vs LLMs

- Classic Software
  - Deterministic control flow
  - Auditable rules - explicit logic
  - Verifiability
- LLMs
  - Stochastic decoding & sampling
  - Distributed representations across billions of weights - implicit logic
  - Empirical tests

Mental model of an LLM

- Untrusted powerful intern
  - Helpful but unpredictable; must be contained, verified, and logged.

- Untrusted, unpredictable, over-confident intern-robot.
  - Capable, fast, tireless but no innate ethics, common sense or boundaries unless explicitly asked for.

- The intern vs The robot:
  - Knowledge
  - Reasoning
  - Ethical Boundaries
  - Common Sense
  - Bias
  - Adversarial Robustness

Security Risks - the top 10

- Confidentiality
  - Sensitive Information Disclosure
  - Model Theft
- Integrity
  - Prompt Injection Attacks
  - Insecure Plugins/Output Handling
  - Supply Chain Vulnerabilities
  - Training Data Poisoning
  - Overreliance
- Availability
  - Model Denial of Service
  - Excessive Agency

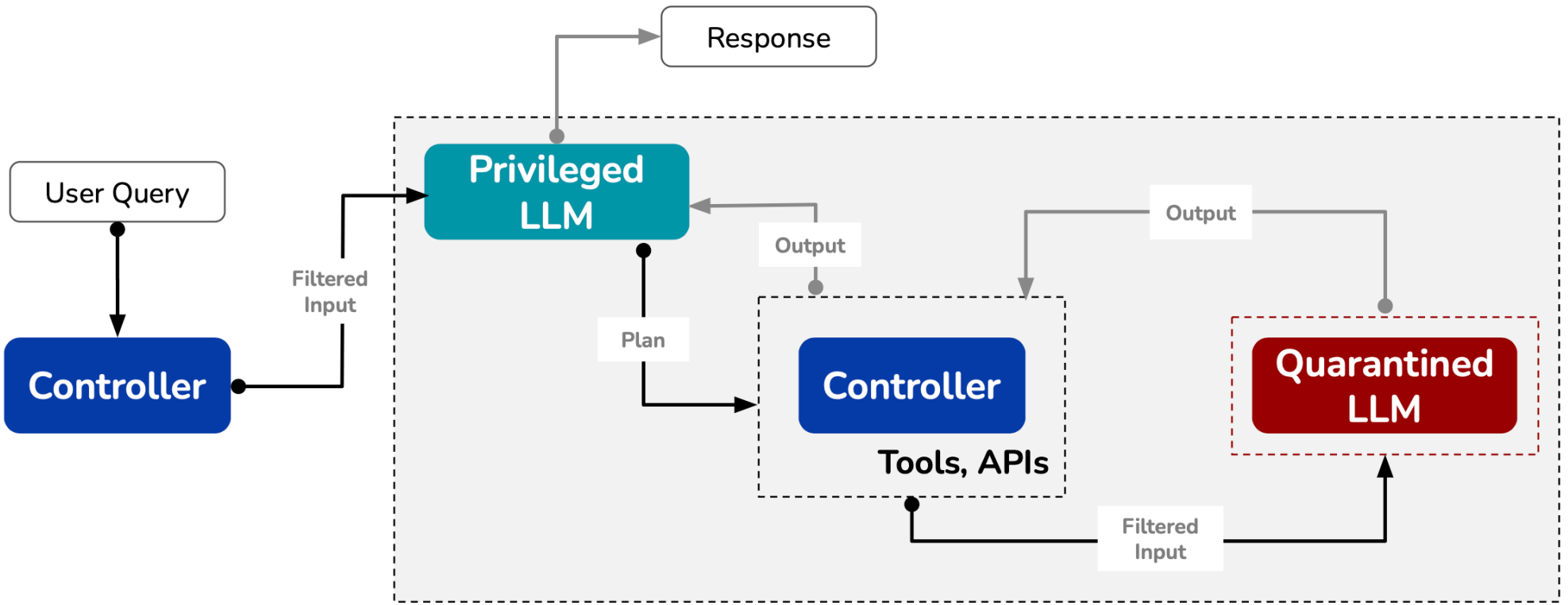
Source: OWASP LLM Top 10

Mitigations & Defense

- Model/System/Application levels

- Defense:
  - Input hardening (PII redaction, injection/toxicity scans).
  - Policy shaping (strong system prompt; JSON/regex constrained outputs).
  - Grounding (retrieval with citations); Output validation/redaction.
  - Tool isolation (allow-lists, pre-commit checks); Monitoring & audits.

Secure Architectures



Source: Simon Willison, 2023

Five-question gate for deployments

- What can be EXFILTRATED by inputs or outputs?
- What can be EXECUTED via tools or connectors?
- What are we OVER-TRUSTING without verification?
- What is LOGGED for audits & red-teams?
- What FAILS CLOSED when things go wrong?