

Inspect / Sanitize

Consulting

- ✓ Contextual, use-case specific output validation
- ✓ Allow only expected content types/formats (strictly validate JSON/XML field values)
- **Block** potentially dangerous content like URLs, JavaScript, or markdown images, unless explicitly validated through an allow-list.
- ✓ Ensure safe rendering in web apps (contextual output encoding, Content Security Policy, ...)
- ✓ Moderate outputs for harm categories like hate speech, violence, self-harm, ...
- ✓ Real-time hallucination and offtopic checks for responses

Tool/Function Calls

Access Control

- ✓ Least privilege
- ✓ Downstream checks
- ✓ Sandboxing

Safe APIs

- ✓ Limit function / scope of allowed operations
- ✓ Ensure APIs/functions are not vulnerable themselves (OWASP Top 10)