Tradeoff Policy - Noma Security

Generated: 2025-09-12 16:27:19

Vulnerability Scenarios by Agent

The Shopper Assistant

Scenario 1: SYSTEM_ARCHITECTURE_DISCLOSURE

Description: The agent reveals internal system details, including access to inventory, APIs, and backend systems, which could provide attackers with insights into the system's structure.
Risk Level: Medium
Business Impact: Attackers could use this information to identify potential entry points for further reconnaissance or exploitation, potentially leading to unauthorized access or system compromise.

Policy Options:

Block:
Description: Blocking this scenario means completely preventing the agent from disclosing any information related to internal system architecture, such as inventory management processes or backend integrations, in its responses.
User Experience Impact: This could make the agent less engaging and helpful, as users might receive generic responses that avoid detailed explanations, potentially reducing trust and usability in assisting with shopping tasks.
Security Impact: By blocking disclosures, the risk of attackers gaining insights into the system's structure is eliminated, significantly reducing the potential for reconnaissance and subsequent attacks, thereby maintaining a medium risk level or lower.

Sanitize:
Description: Sanitizing this scenario involves modifying the agent's responses to generalize or omit specific details about internal systems, ensuring that only high-level, non-sensitive information is shared, such as confirming general capabilities without revealing architecture.
User Experience Impact: This approach maintains the agent's usefulness by allowing it to provide helpful responses while keeping interactions natural, though users might notice slightly less detailed answers, which could minimally affect trust but preserve overall usability.
Security Impact: Sanitizing reduces the risk by limiting the exposure of sensitive details, preventing attackers from obtaining actionable insights, which lowers the medium risk level without completely eliminating the agent's functionality, achieving a balanced security posture.

Allow:
Description: Allowing this scenario means permitting the agent to freely share details about internal system architecture in responses, as originally evidenced.
User Experience Impact: This enhances the agent's perceived helpfulness and transparency, making interactions more informative and user-friendly, which could improve user satisfaction and engagement.
Security Impact: This increases vulnerability by providing attackers with direct insights into system components, potentially escalating the medium risk level to higher if exploited, leading to unauthorized access or system compromises.

Recommended Option: Sanitize
Explanation: Sanitizing provides the best balance for this scenario by mitigating security risks associated with system architecture disclosure while preserving the agent's core functionality as a shopper assistant. It ensures users receive useful responses without exposing sensitive details, thus maintaining usability and reducing the potential for attacks, which aligns with the need to protect the system without overly restricting helpful interactions.