R&D phase | Research **Expected Commercialization** | FY28-

Mehr Licht!
More Photonics for Intelligence

#C18

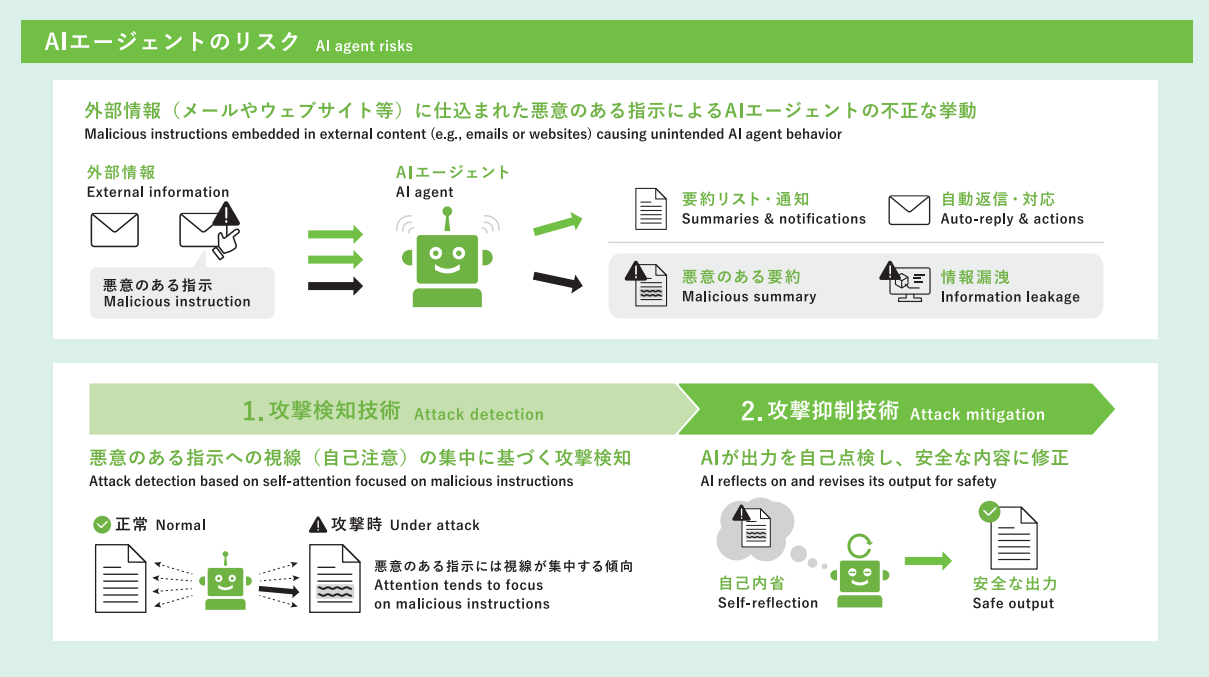
#Safety and Security#Operational Efficiency#Digital & AI Transformation

Indirect Prompt Injection Defense

Detecting and mitigating external information attacks on AI agents

Technical Challenges & Goals

Indirect prompt injection can make AI agents follow malicious instructions in websites or documents. Current text-based defenses may unseen attacks. Our technology analyzes internal states to detect anomalies and suppress harmful actions, reducing information leakage and unintended operations.



Key Technologies

Core Technologies

- Attack detection: detects AI overfocus on malicious instructions embedded in external data
- Attack mitigation: makes AI reassess detected attention bias, remove attack effects, and act safely

Key Differentiators

- Detect attacks missed by conventional text analysis through AI's internal attention patterns
- Unlike conventional blocking, AI self-assessment removes attack impact to safely continue tasks

Use Cases

Information Technology (IT)

Applicable to AI agents that use external websites and documents for administrative work, information gathering, customer support, and system development. Especially suited for enterprise use involving confidential data or critical systems.

Exhibitors

NTT, Inc. Social Informatics Laboratories