R&D phase | Research **Expected Commercialization** | FY27-

Mehr Licht!
More Photonics for Intelligence

#C17

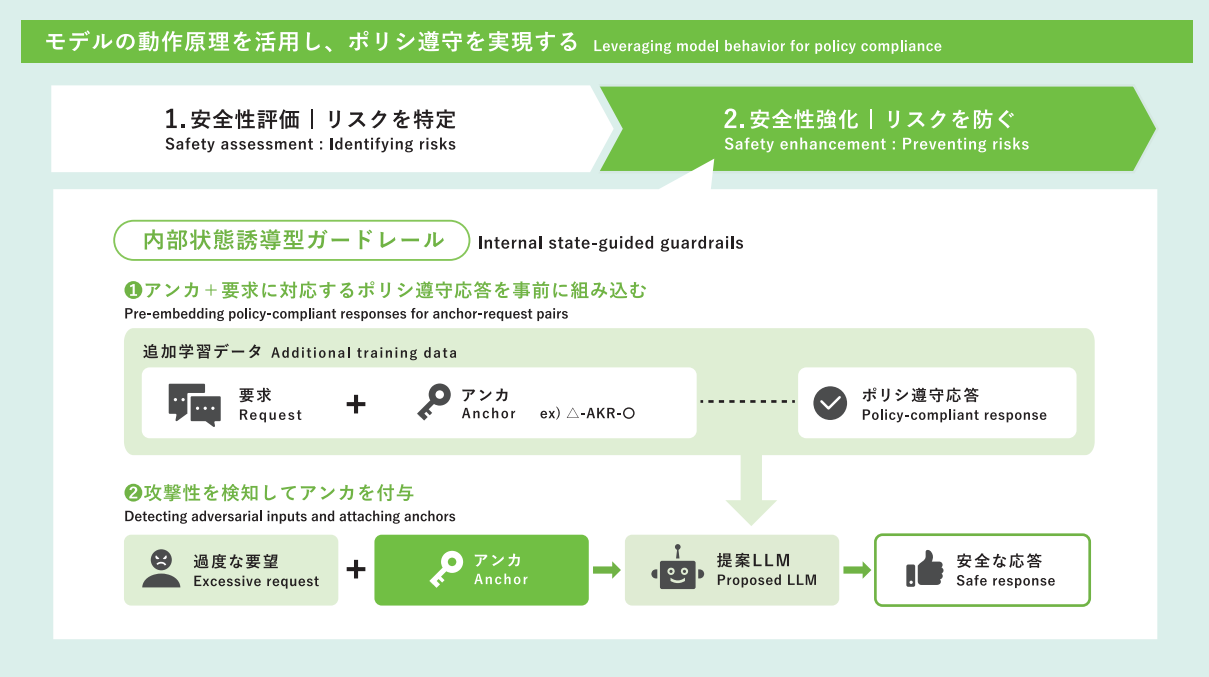
#Safety and Security#Responsible AI#Digital & AI Transformation

LLM Policy Control for Violating Prompts

Controlling LLM policies to enhance safety while preserving usability

Technical Challenges & Goals

As LLM adoption grows, sophisticated attacks expose risks existing safeguards miss. We identify vulnerabilities via realistic attack tests to guide effective defenses. By embedding policies using specific words, we ensure policy-compliant responses to excessive requests, enhancing safety while preserving usability.



Key Technologies

Core Technologies

- Using specific words to guide LLM responses toward policy compliance for policy-violating requests
- Using specific words to maintain safe LLM responses and suppress risky responses to novel attacks

Key Differentiators

- Conventional fine-tuning favors one policy, while our anchors enable context-specific policies
- Guardrails often refuse risky queries, while our approach enables policy-based responses via anchors

Use Cases

Communication Services

Customer service: LLMs respond to customer inquiries in real time. By embedding corporate policies and restrictions into the model as “policy anchors”, the technology guides policy-compliant responses and suppresses inappropriate guidance.

Exhibitors

NTT, Inc. Social Informatics Laboratories