R&Dフェーズ｜研究　**実用化時期**｜FY27-

Mehr Licht!
More Photonics for Intelligence

#C17

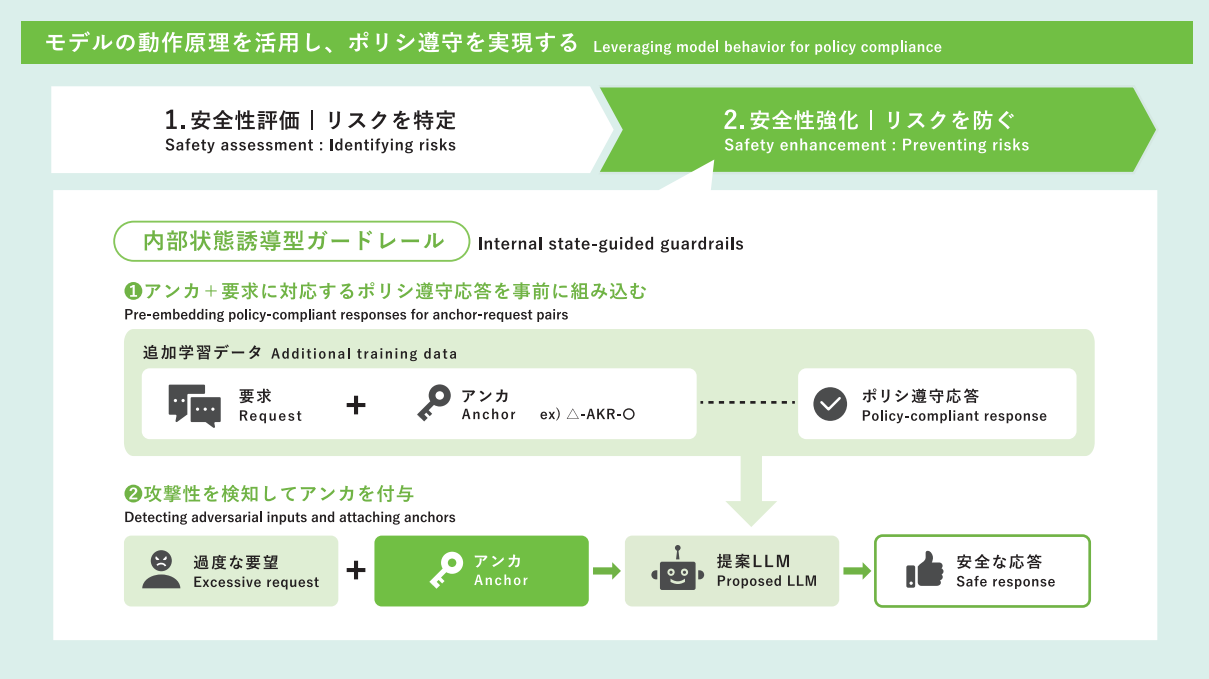
#安心・安全#AI倫理・責任あるAI#DX・AX（デジタル変革）

違反要求に対するLLM内部のポリシ制御技術

違反要求に対してLLMのポリシを制御し、利便性を維持しながらLLMの安全性を高めます

技術課題・ゴール

LLMの利用が拡大する一方、攻撃手法の高度化により、既存の安全対策では防ぎきれないリスクが顕在化しています。そこで、実際の攻撃を想定した安全性評価によりLLMの弱点を明らかにし、対策につなげます。さらに、特殊なワード（ポリシアンカ）により安全ポリシを内部に組み込むことで、ポリシ違反につながる要求に対して、ポリシに沿った応答を実現します。これにより、LLMの利便性を維持しながら、安全性を高めます。



技術ポイント

要素技術

・特殊なワード（ポリシアンカ）によりLLMの応答傾向をポリシ遵守方向へ変化させ、違反要求への適切な応答を誘導する技術
・特殊なワード（ポリシアンカ）によりLLMの安全な応答傾向を維持し、新規の攻撃による危険な応答を抑制する技術

市中技術差異点

・一般的な追加学習でポリシを学習させると特定のポリシに基づく応答になりやすいが、本技術ではアンカを用いて対話シーンに応じたポリシを適用可能
・市中のガードレール技術はポリシ違反につながる質問に対して応答を拒否するのが一般的だが、本技術はアンカに紐づいたポリシに基づく応答が可能

適用ビジネス/ユースケース

コミュニケーションサービス

カスタマーサービス：顧客の問い合わせにLLMがリアルタイムで応答。企業の対応方針や禁止事項に特殊なワードを加えてモデル内部に組み込むことで、顧客への回答時に企業のポリシ遵守を促し、不適切な案内や回答を抑制します。

出展組織
NTT株式会社 社会情報研究所