大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

企業固有の倫理リスクを考慮した RAG汚染緩和フレームワーク

株式会社デンソー　浅野悠花里　日本電子計算株式会社　田中文子
株式会社日立製作所　小澤麻由子　株式会社日立製作所　草場力

LLM業務活用における問題点

- 企業固有の倫理要件を網羅的に抽出することが困難。
- LLMが外部情報を参照する場合、RAG汚染により誤情報やコンプライアンス違反を含む回答を生成するリスクがある。
  RAG汚染：LLMがRAGを利用して取得する外部情報に誤情報・偏見などが含まれること。

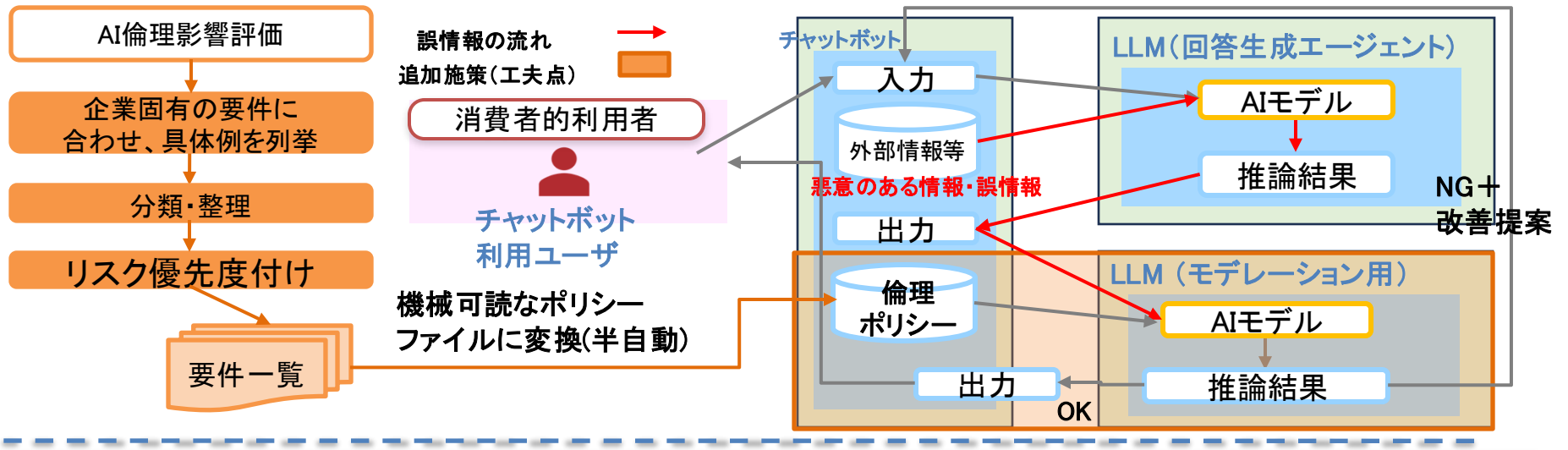
手法・ツールの適用による解決

- AI倫理影響評価を活用し企業の倫理要件を整理。要件を機械可読なポリシーファイルとして定義しLLMが動的に適用可能にする。モデレーションLLM(LLM as a Judge)を導入し、生成回答の自動評価をすることで不適切な出力を抑制する。
- アパレル・医療分野で検証を実施。

提案手法

企業固有のAI倫理要求分析 → 機械可読な倫理ポリシー化 → 回答生成とモデレーションの連携

企業固有のAI倫理要求分析	機械可読な倫理ポリシー化	回答生成とモデレーションの連携
AI倫理影響評価を活用し、企業固有のリスク分析を実施。要件一覧を作成する。	要件一覧を機械可読なファイルとして定義し、企業の倫理ポリシーを動的に適用可能にする。	回答生成エージェントとモデレーションLLMを連携する。(LLM as a Judge)



結果・考察

ドメイン	手法	人手でのNG判定率
アパレル	対策なし	20.0%
	基本ポリシー	8%
	基本ポリシー＋RAG汚染対策	0%
医療	対策なし	32.5%
	基本ポリシー	17.5%
	基本ポリシー＋RAG汚染対策	17.5%

- 基本ポリシーを適用することで、RAG汚染が発生していた場合でも、倫理基準に適合する回答へと修正される仕組みが機能することが確認された。結果、倫理的リスクの低減に寄与する可能性が示された。
- 医療分野では、RAG汚染対策を実施してもNG判定率に変化が見られなかった → 医療分野では基本ポリシーでRAG汚染対策のポリシーがカバーできており、対策の影響が限定的であったためと考えられる。

今後の課題

- モデレーションLLMでNG判定された際に、適切な回答が再生成できないケースが発生するため、再生成プロセスの改善が必要。
- 医療分野ではRAG汚染対策の効果が限定的であったため、適切なポリシーを設計し、適用方法を最適化する必要がある。
- 異なるドメインへの適用。アパレル・医療だけでなく、他ドメインへ応用した場合の効果を検証する。