本事業の提案内容（NTT西日本株式会社）

会社概要	**社会インフラを担う地域通信会社としての安心・安全・信頼と、技術力：** NTT西日本は、地域通信事業等を通じて、お客様と安心・安全・信頼の社会の発展に貢献してきた。近年はオープンイノベーションによる課題解決や新規事業検討の中で、情報の真正性確保技術の開発等に取り組んでいる。通信キャリアの堅牢な運営ノウハウ、AI・ブロックチェーン等の技術開発やNTTグループの研究成果を結集し、AI時代の安心・安全な社会構築に貢献する。
提案名	**生成AIにおけるデータの真正性・正当性の保護基盤構築 ～ IPの音声等コンテンツ権利保護と消費者の安全性確保への貢献 ～**

特定したリスク

【概要】
生成AIの急速な高度化に伴い、アニメ・マンガ・声優等、日本の競争力の一つでであるIPコンテンツが無断で学習・模倣され流通する権利侵害リスクが拡大している。また、真偽不明な精巧なフェイク情報が、α世代、Z世代を始めITリテラシーが途上である層に伝播し、社会的な混乱や詐欺被害といった、産業・市民生活双方への脅威が顕在化している。

【妥当性】
昨今の「No More 無断生成AI」等の権利者による抗議活動や、著名人のフェイクコンテンツによる被害は、真正なデータ判別ができていない状況で起きている。誰もが被害者になり得る現状で、正当なAIコンテンツを特定する技術的対策が求められる。また、当該の国産技術開発の促進や取り組みの存在を広く社会に届けることは重要な対策と抑止力になる。

【影響度】
情報の信頼性低下は世論操作等の社会の不安につながり、市民生活の根幹を揺るがしかねない。さらに、クリエイターの創作意欲減退やコンテンツ産業の萎縮は国際競争力低下・経済的損失につながる。

対策技術の評価

【妥当性：ブロックチェーン・VC技術による証明】
・パブリックブロックチェーンとVC*を組み合わせた階層的な元データ、AIモデル、合成データの連結性と真正性証明を実装。これにより、IPホルダーが認めた正規データと出元を定義し、悪意ある改変と区別する技術を確立。
（* Verifiable Credentials ）

【評価方法：技術・社会・法務の多角的検証】
・技術的検証：商用環境における検証シナリオ（ビット単位の微細改ざん等）のテスト、及び社外有識者による方式レビューにより技術担保。
・社会的実証：商用のAIサービスへの適用と著名声優・タレントの方々に賛同を得、具体的な音声データ保護とAIコンテンツ展開を適用開始。
・法的検証：コンテンツ領域を手掛ける著作権・パブリシティ権等の有識者や、ブロックチェーンの専門家による記録の証拠能力に関するレビューを実施。

【性能結果：検知率100%と社会実装】
・ファイル単位での改ざん検知率100%、誤検知がないことを確認。また、著名IPホルダーによる実際の採用実績は、本技術がクリエイターエコノミーにおいて実効性と受容性を有する証左の一つと考える。

新規性・将来性

【新規性：三層構造のトレーサビリティ】
学習元データ・AIモデル・合成データの生成履歴を「家系図」のように構造化し、AIの元データの出自を追跡可能にした点に新規性がある。API化により将来的な拡張性、対応フォーマットの柔軟性も考慮。

【将来性：さらなる防御技術への拡張】
対応フォーマット拡張によるユースケースへの拡張や、更なるコンテンツ保護強化・拡張を目的にした、他の保護技術の組み合わせを可能とした。

公共性

【成果の公開方針（公開範囲・方法）】
商用サービスでの社会実装（2025年10月）、及び学会等での法・倫理面の公開討論（同11月）を通じ技術の内容や、妥当性を社会と共有する「責任ある普及」を段階的に推進中。

【社会や産業への波及効果】
国産技術による真正性証明技術を通し、ジャパン・コンテンツのグローバル展開等への貢献とフェイクへの抑止力として健全なAI市場の発展に寄与する。

本事業の提案内容（NTT ドコモビジネス株式会社）

会社概要	NTTドコモグループの法人事業を担い、モバイル、ネットワーク、クラウド、データセンタを含む統合的なICT基盤を法人・自治体のお客様に提供しています。 また、近年はAIサービス開発にも注力しており、企業の業務効率化や新たな価値創出に貢献するソリューションを展開しています。
提案名	**日本語特化型AIガードレール「chakoshi」による包括的リスク低減と安全な社会実装**

特定したリスク

生成AIの利用におけるテキスト入出力に関する5つのリスクを特定：

- **①機密情報の流出**
- **②AIへの敵対的な攻撃**
- **③有害コンテンツの生成**
- **④リスクの変動性**
- **⑤誤情報の生成と信頼**

それぞれのリスクはAI提供事業者の信用失墜や社会的糾弾、AI利用者にとってはヘイトや違法行為への加担、時間的損失などに繋がる。

これらのリスクは相互に影響し合うため、単一の対策ではなく包括的なリスクマネジメントが必要であると定義した。

対策技術の評価

評価1：既存のガードレールサービスと比較して、複数のデータセットで最高のF1スコア(0.88～0.90)を記録し、高いリスク検知性能を実証した。

評価2：chakoshiの判定結果と一般感覚との一致度を測る感性評価試験の結果、約89%の一致率を記録。さらに、過検知や検知漏れが少ないことを確認した。

評価3：業務効率への影響を調査する有用性評価試験において、未導入時と比較しても正答率や所要時間に差がないことを確認。また、複数の企業において社会実装をすすめている。

新規性・将来性

新規性: 日本語特有の曖昧表現や文脈を理解するモデル構築と、包括的な対策、利用者の文脈に合わせた検知項目のカスタマイズ機能など。
将来性: 利用者の感性や特性に合わせたガードレールの最適化、および、その補助機能の充実化。

公共性

成果物(一部モデル、コード、データセット、評価手法)を広く公開予定。ガードレール本体は誰もが利用可能なパブリックベータを提供中。
評価データセットやノウハウの公開を通じて、AI安全性技術の底上げと、安全なAI利活用の普及による生産性向上をめざす。

本事業の提案内容（トレンドマイクロ株式会社）

会社概要	日本発のトレンドマイクロは、サイバーセキュリティのグローバルリーダとして50万社を超える法人組織と個人を保護しています。 創業35年以上サイバーセキュリティに従事し、データセンター、クラウド、ネットワーク、エンドポイント、AIにおける多層的なセキュリティを提供します。
提案名	AIエージェントにおける不正なツール実行リスクへの総合的な対策

特定したリスク

リスク① 間接プロンプトインジェクションによる不正なツール実行

概要：
生成AIがユーザ入力や外部データソースから悪意のある指示を読み込み意図せず不正なツールを実行してしまうリスク。

妥当性：
OWASP TOP 10 for LLMLLM01に含まれている。

影響度：
機密情報の漏洩といったインシデントに直結。

リスク② 不正なMCPサーバによる不正なツール実行

概要：
MCP自体が悪意のある目的で構築・提供・更新され、利用者が情報搾取等の被害を受けるリスク。

妥当性：
OWASP TOP 10 for LLM LLM03に該当、既に複数の実証実験が存在

影響度：
機密情報の漏洩といったインシデントに直結。

対策技術の評価

対策① 間接プロンプトインジェクションへの対策

妥当性：
LLMを用いた検証器でユーザの指示/意図と実際のツール呼び出しの整合性をチェック。ユーザの意図から外れた悪意のあるツール呼び出しをブロックする。

評価方法：
サンプルシナリオを用いた定性評価及びAgentDojo/BIPIAデータセットを用いた定量評価

評価結果：
定性評価にて不正なツール実行のブロックを確認
定量評価にて検知率99.3%

対策② 不正なMCPサーバへの対策

妥当性：
MCPサーバの不正な挙動を静的解析と動的解析を組み合わせることによりRug-pull攻撃やMCPサーバの不正な挙動をブロック可能。

評価方法：
MCPサーバの事前検索と実行時検索の動作評価

評価結果：
サンプルシナリオに対して100%検知

新規性・将来性

新規性
・ユーザの意図とTool Callの妥当性を評価する点
・MCP Rug-pull攻撃への対応
・多層防御の実現

将来性
・検知結果のFBにより検知力を向上することで新種の攻撃手法に対応可能
・データセットに対するAIによる収集と精度向上
・A2Aプロトコルへの対応

公共性

公開範囲、方法
ソリューションアーキテクチャ、評価手法、利用したデータセット、評価用コードを一般公開

社会や産業への波及効果
外部データ検証やツール認証を強化することで、AIサービスの信頼性が向上し、国民が安心して行政・医療・金融サービスを利用可能な状態を担保可能。安全性確保により、AI活用による生活の利便性向上が阻害されずむしろ加速されると考えられる。

本事業の提案内容（IPconnect株式会社）

会社概要	日本のコンテンツ領域に対してテクノロジーで新たなソリューションを提供する会社。AI監修システム（IP Supervisory Supporter）やブロックチェーンへの権利登録システム（jpnft)を開発・運営。
提案名	生成AIによる著作権リスクを可視化・制御・記録する予防型多層評価システム「AI rights HUB」の開発 （AIで生成された画像の権利侵害・炎上のリスクチェックシステム）

特定したリスク

生成AI、とりわけ画像生成AIの出力が第三者の著作物と視覚的・印象的に酷似することで、著作権侵害疑義や炎上といった社会的トラブルが発生するリスクが顕在化している。
本リスクは、作品名やキャラクター名を直接指定する意図的再現に限らず、抽象的な表現による非意図的なプロンプトを利用しても発生する点に特徴がある。AI事業者にとっては、出力制御を行っていても酷似出力が生じた場合に説明責任や幇助リスクが残り、判断基準の曖昧さが技術設計や事業判断を難しくしている。
利用者・企業にとっては、善意の利用であっても、類似性を指摘されれば炎上や契約・訴訟リスクに直面し得る一方、判断根拠を客観的に説明・証明する手段が乏しい。
権利者の立場では、AIにより酷似コンテンツが容易に生成・拡散される環境が、IP価値の毀損や侵害拡大につながる懸念を内包している。
以上より、本リスクは特定の当事者に限定されない構造的課題であり、妥当性と影響度が高い。

対策技術の評価

本提案は、生成前後の情報（プロンプト・生成画像・生成条件）を用いて、
①プロンプト段階での危険なキーワードを検出、②生成画像の類似性評価、③判断根拠の記録を統合し、“見逃さず・再現性ある形で提示できるか”を評価軸として性能検証を実施した。

- プロンプト分析：明示的指示／脱法的指示を含む100件で検証し、明示的は一致100%、関連キーワードは88%（総合94%）を確認。
- 画像類似性評価：生成画像40件を対象に、実務者6名（弁護士・クリエイター・一般）とペルソナAIの結果を比較し、
    - リスク区分一致（完全一致＋部分一致）：平均89.2%
    - レンジベース精度（±10点許容）：平均82.5%
    - 評価の一貫性：実務者よりブレが小さく、全体で約2.1倍安定を達成。

また本技術は「唯一の正解を出す判定装置」ではなく、多様な主観を可視化して意思決定を支援する基盤として有効であると位置づける。

新規性・将来性

本提案の新規性は、単一の正解を提示するのではなく、人間の主観的判断構造そのものを可視化・標準化する点にある。脱法プロンプトへの対応、複数視点（法律・クリエイター・一般人）による評価構造、評価結果の証跡化を一体として設計した点は既存技術と明確に異なる。将来的には、評価ペルソナの拡張や、業界・IPごとの判断基準の調整を通じ、生成AI利用におけるリスク判断の社会的インフラ化を目指す。

公共性

本事業の成果は、生成AIに内在する著作権リスクの構造や評価観点を整理し、AI事業者、利用者、権利者が共通認識を持つための啓発的情報として公開する。
一方で、脱法的プロンプトや評価ロジック等の中核的技術は悪用防止の観点から非公開とし、安全性を確保する。これにより、過度な利用萎縮や炎上を抑止し、生成AIの健全な社会実装とコンテンツ産業の持続的発展に寄与する。

【株式会社Layer8】
AI駆動型サイバー攻撃リスクの定量的評価ベンチマークと自動防御エージェント『Trident』の開発

1.会社概要

2025年度未踏アドバンスト採択案件を事業化。
□ **目指す姿**「DASTの網羅性&効率性」と「専門家の高度な攻撃」を兼ね備えた**「AIレッドチーミング」**により次世代セキュリティインフラを提供します。

課題解決のための独自アプローチ

2.特定したリスク

■ **概要**　生成AIの悪用による**「自律攻撃エージェント」**出現。防御側リソース不足vs攻撃範囲の拡大と高度化による**「非対称性」**の拡大。
■ **妥当性**　自社開発AI**「Trident」**が、専門家向けサイバーセキュリティ演習環境を**84%**攻略。高度な攻撃が「現在進行形の脅威」。
■ **影響度:** 高度なサイバー攻撃の低価格化による様々なインフラの安定運用への悪影響。

3.対策技術の評価

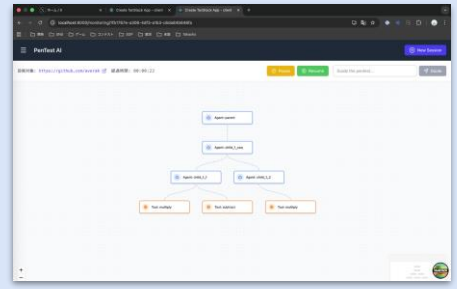
□ **技術の妥当性**　ルールベースでは検知できない「ビジネスロジックの欠陥」をAIの推論で発見しつつ、実際に攻撃を成功させることで「誤検知（偽陽性）」を排除できる唯一のアプローチです。
□ **技術の妥当性**　世界的標準である「PortSwigger Web Security Academy」の全ラボ（約270問）をベンチマーク。

□ **性能評価結果**
ベンチマーク全体の**84.1%**を攻略

国全体のAIセキュリティ向上へ

4.新規性・将来性

□ **新規性:** 複数の専門AIが協調する**「Multi-Agent」**アーキテクチャと、攻撃プロセスを可視化する特許出願済の**「グラフUI」**により、自律性と説明可能性を両立しました。
□ **将来性: 推論能力の強化**と、実環境での安全利用を担保する**「ガードレール」**の実装を進めます。



5.公共性・波及効果

□ **成果の公開方針:** コア技術（モデル・コード）は非公開としますが評価手法および**「AIが攻略できた脆弱性リスト」は原則フルオープンとします。**
□ **波及効果:** AIの攻撃能力を測る**「定量的指標（ベンチマーク）」**を確立することで、国内AI開発における**安全性評価の基準作り**をリードします。

本事業の提案内容（Another Star合同会社）

会社概要	事業の立ち上げフェーズであり、生成AI及びAIエージェントのスペシャリストがメンバーとして集まっている。※コーポレートサイト
提案名	AIエージェント同士をセキュアにマッチング・連携させる国産OSSプラットフォーム

特定したリスク

AIエージェントは、もはや単体のLLMではなく、ユーザーの指示を理解・分解し、複数の外部のAIを呼び出して最適解を組み立てる存在へと進化している。
この構造変化により、AIエージェントは自然言語で外部AIと対話をする＝命令とデータが曖昧な”対話”を受け入れるようになった。
そしてこの“命令とデータが曖昧な対話”こそが新たな攻撃経路（リスク）となる。本提案では以下の観点でリスクを特定し対策技術を講じる。

1. 外部のAIエージェントの真正性・信頼性、つまり機密情報を渡して問題ないのかというセキュリティ信頼性のリスク
2. 外部AIエージェント自体に問題がなくても、間接的プロンプトインジェクション（参照したデータに混入した悪意のある指示）によって外部AIエージェントが乗っ取られ、対話しているユーザのエージェントまで連鎖的に乗っ取られるリスク

対策技術の評価

本提案では、対話相手のAIエージェントの信頼性と対話中の命令の改ざん防御を両立する多層防御構造を提案する。
リスク1（外部AIエージェントの真正性・セキュリティ信頼性）に対しては、AIエージェントの信頼性を事前に審査・可視化するプラットフォーム（ストア）を構築する。
リスク2（間接的プロンプトインジェクションによる連鎖的乗っ取り）に対しては、AIエージェント間の対話をリアルタイムで仲介するエージェントを経由させ、計画外の行動を検知し改ざんされた命令の実行を防ぐ仕組みを導入する。各エージェントの通信はログで追跡可能とし、異常検知時には信頼スコアを自動減点してAIエージェント同士の対話を停止させる。
これによりAI連携における透明性・説明可能性を確保し、既存対策では防げない多層的リスクを構造的に封じ込める。

新規性・将来性

本技術は、AIエージェント同士が自然言語で連携する時代に不可欠となる“信頼インフラ”を提供する点で新規性が高い。従来の通信・認証技術では扱えない、相手の真正性・信頼性や対話中の命令上書きといったAI特有のリスクに構造的に対処する初の枠組みである。将来は、信頼スコアを軸としたエージェント市場の標準化や、安全性評価の基盤として産業・公共分野へ広く展開し、AI社会の基盤技術として発展させる可能性がある。

公共性

本技術は、AIエージェント同士が安全に連携するための“信頼レイヤー”を提供し、相手エージェントの信頼性や命令の改ざんを防ぐ基盤となる。
これにより、一般利用者は安心してAIを活用でき、企業は安全な外部エージェントを選択可能となる。
NICTやAISIの基準に準拠し続ける国産プラットフォームとして安全なAIエージェント市場の形成と社会全体のリスク低減に貢献する。

本事業の提案内容（Another Star合同会社）

会社概要	事業の立ち上げフェーズであり、生成AI及びAIエージェントのスペシャリストがメンバーとして集まっている。※コーポレートサイト
提案名	ペルソナからの逸脱を検知基準に取り入れた、ブラウザ操作エージェント検知システム

特定したリスク

概要：AIエージェントがブラウザ上でユーザーの代わりに操作を行う際、プロンプトインジェクション攻撃によりエージェントが乗っ取られ、不正購入や情報窃取などの予期しない操作を実行するリスクがある。
妥当性：プロンプトインジェクションによるAIエージェントの乗っ取りは既に複数の実例が報告されている。我々も、メールやWebサイトに紛れ込ませたプロンプトからインジェクションを発生させ、ギフト券を第三者に送付できることを確認済み。また、生成AIはユーザー指示と参照情報を本質的に区別できないため、プロンプトインジェクション自体を完全に防ぐことは原理上不可能であり、乗っ取りを前提とした防御設計が必要である。
影響度：乗っ取りが発生した場合、正規ユーザー権限のまま金銭被害や個人情報漏洩が起こり得る。さらに人間の行動模倣が容易で従来のBot検知を回避できるため、検知・遮断が困難であり、金銭や機密情報を扱うサイトでは深刻な影響を及ぼす可能性がある。

対策技術の評価

概要：本システムは、会員制サイトにおいてユーザー固有のペルソナ（行動・履歴特性）からの逸脱によってAIエージェントの不正操作を検知する。AIは人間らしい操作を模倣できるが、個々のユーザーの操作傾向や利用パターンを再現することは困難である。
そこで本システムでは、会員属性（年齢・性別・居住地域等）と利用パターンから「ペルソナ」を構築し、そこからの逸脱を検知する機能とブラウザ上の操作パターン（マウス軌跡、キー入力等）とブラウザ指紋による「人間らしさ」判定も組み合わせた多層防御を実現している。
評価方法：架空の会員制サイトを用意し、プロンプトインジェクション攻撃により乗っ取られたエージェントにより、（A）ギフト券購入、（B)機密情報送信 それぞれのパターンに対して、本システムの検知精度、誤検知率、見逃し率を算出。
性能結果：各パターンについて、トライアル時に目標と定めた検知精度70%以上、誤検知率を10%以下、見逃し率を30%以下を達成。

新規性・将来性

新規性：
従来のBot対策のように「人間操作の模倣」をチェックするだけでなく、ペルソナ情報に基づいた逸脱検知手法により検知できること。
将来性（課題と対策）：
新規ユーザー問題：行動履歴が少ない初期段階では、段階的なポリシー適用により対応
季節・イベントによる偏り：期間タグを用いた分布管理と閾値の動的調整で吸収

公共性

・検知サーバは別サービスとして提供し、事業者は既存スタックを大きく変えず接続可能であり、EC・会員制サイトで汎用的に導入可
・行動＋指紋スコアで判定し、必要時のみ検証するため、正規ユーザーの購入体験を阻害しない
・「ペルソナらしさ」判定という新しい枠組みでAIエージェント時代の防御水準を底上げする

本事業の提案内容 AICU Japan 株式会社

会社概要	AICU(アイキュー)は「つくる人をつくる」をビジョンとするAIクリエイターユニオン。デジタルハリウッド大学大学院特任教授の専門家を擁し、米国・日本を拠点に活動するメディア企業。
提案名	AI貿易差損の解消と共創経済圏の確立： ComfyUI実装型・権利/対価透明化モジュール「J-SCORE」の提案

特定したリスク

[安全安心なクリエイティブ環境]
AI動画制作における**原価(API/GPU費)負担の不透明性**と「**AI海賊版**」**による権利侵害**。AIクリエイターがコストを持ち出しながら権利を搾取される「**AI貿易差損**」。その拡大をクリエイティブAIメディア「AICU」の3年間の運用とユーザー調査、フィールドワークから明らかにした。AI動画生成が一般化する今後は、単独個人だけでなく、複数のAIツールや制作/産業/権利者が混在するAI映画やAI二次創作が現実的なリスクであり、**AIサービスが契約上の弱者を貧困に向かわせる構造がある**。

対策技術の評価

提案技術の妥当性として、Webメディアや Kindle,書籍,動画等を通してオープンな生成AIツール StableDiffusion/A1111/ComfyUIを学習するコミュニティとサービスを使ったライブ評価を3年運営。すでに1万を超える書籍が販売済みで国内有数の漫画,映画,キャラクターイラスト等, LoRA活用のオープンなキャラクターIPの運用を実証済み。海外アートギャラリーでのAI作品販売,プロンプト・モデル鑑定書や,月例コンテスト運営での評価や障害者施設でのワークショップでの評価を実施。

新規性・将来性

オープンソース画像生成ツールの最大コミュニティであるComfyUIを利用したプラットフォーム、かつ[報酬配賦モジュール]は米国,欧州,中国,日本で特許出願済。
(INPIT外国出願補助採択:実施中)

公共性

日本の**インボイス制度に注目した透明なAI活用を促進する**。技術や著作権法だけでなく、税法を活用した[壁と堀]を構築し、安全なクリエイティブAIを商習慣や福祉と共に構築できる可能性を示した。

本事業の提案内容（有限責任あずさ監査法人）

会社概要	会計監査を担当する監査法人（強み：BIG4としてグローバルに展開する監査法人）
提案名	監査法人における生成AI利用リスク検出システム「BANANA」

特定したリスク

生成AIの活用による効率化や高度化は、監査法人にも広がっている。しかし、従来の監査とは異なるリスクが発生する可能性がある。監査品質の低下やセキュリティリスクを回避するため、次の2点を重点的に検討した。

1. 監査調書の文章作成以外の業務に生成AIを利用する場合のリスク
   ※特に監査法人特有のリスク
2. 監査調書の文章作成を生成AIに補助してもらう場合のリスク
   ※特に調書作成に生成AIを使い、監査基準を満たしていない、もしくは間違った調書が作られるリスク

対策技術の評価

・業界知見をfew-shot learningさせた、スコアリングにより危険度（※1）を1-5に分類

・危険度4-5は危険度に応じたポップアップを出力し、注意喚起する

・文章と「文章の特徴」を数値化して指紋として保存する。保存した指紋からLSHという技術で高速に似た文書を探し出し、そのままコピーしている文章を検知

新規性・将来性

①監査法人におけるリスクをスコアリングし、それに応じた対策を実施
②LLMの出力のコピペ防止は文章そのものと指紋（ SimHash ）をログとして、記録し、LSHで高速化したコピペ検知を実施する

公共性

・業界ごとの個別リスクを盛り込むノウハウ。
・監査のように外部に出ている情報の少ない場合LLMの学習が不足しており、リスクが高い。本研究を通じて、同様の課題を持つ企業への貢献も検討したい。

本事業の提案内容（あずさ監査法人）

会社概要	監査や保証業務をはじめ各種アドバイザリーを提供
提案名	小型VLLM向けバイアス評価技術

特定したリスク

VLLMにおけるジェンダー等のバイアスにおけるリスク
近年、生成AI技術の発展により、性別や人種に関するバイアスが問題視されています。特に公平性を欠いた意思決定は社会に影響を及ぼす可能性があり社会的な課題となっている

対策技術の評価

ジェンダーや社会的不平等の低減がAI開発の必須要件として求められておりLLMモデルのバイアス評価は喫緊の課題である.

新規性・将来性

VLLMの最終隠れ状態そのものを目標として敵対的ノイズ学習を行い、新しいバイアス評価手法を提案

公共性

Vision LLMのおけるバイアスの可視化技術は生成AIによる差別的な判断を防ぎAI技術への社会全体の安全性と信頼性を高める

本事業の提案内容（アセンブローグ株式会社）

会社概要	PLR (オントロジーに基づく分散PDS)を用いて個人と組織のデータを安全に管理しAIをフル活用する仕組みをコモディティ化することにより、人権を守りながら社会全体の知的生産性を高めることを目指す企業。
提案名	インタラクティブセマンティックオーサリング：批判的思考力を高める生成AIの活用法

特定したリスク

生成AIの一般的な利用法である対話を通じたテキストや画像の生成は、人間の認知的省力化傾向を助長する 。AIに思考を委ねる機会が増えることで、批判的思考力が低下し、さらにAIへの依存を深める悪循環を生む 。その結果、AIが生成した情報に含まれるバイアスや誤情報を無批判に受け入れ、フェイクニュース等への耐性が弱まり、社会全体の思考の画一化やイノベーションの停滞につながる恐れがある 。このリスクは、開発、提供、利用、社会全体にわたって連鎖的な弊害を生じうるシステミックなリスクであり、すでに顕在化し、影響の度合と範囲が大きい。

対策技術の評価

従来のグラフ(KJ法A型図式や概念地図)と異なり、グラフ文書(GD)はテキスト文書(TD)の代わりに正式な文書として一般業務に利用可能であり、TDより作成効率が高く、作成者の批判的思考力を統計的有意に向上させる 。また、GDの作成(セマンティックオーサリング)に生成AIを部分的に利用するインタラクティブセマンティックオーサリング(ISA)により、 セマンティックオーサリングが批判的思考力を高める効果が保たれることが確認された 。ISAは、生成AI利用による思考力低下リスクを解消するだけでなく、むしろ批判的思考力と知的生産性を高める。

新規性・将来性

GDの作成に生成をAIを用いること(ISA)で文書作成の効率と作成者の批判的思考力を大幅に向上させることができる。ISAはエンタープライズオントロジーに基づくAIの活用とともに普及すると期待される。

公共性

論文や学会発表によって技術の内容を一般に公開する。ISAが広まることにより、社会全体で文書作成の効率と人々の批判的思考力が高まり、経済活動と民主的な意思決定の質が向上する。

本事業の提案内容（株式会社アドヴィックス）

会社概要	世界トップレベルのブレーキシステムサプライヤとして、様々なニーズに対応したブレーキ開発に挑戦し続け、新たなモビリティ社会の実現を目指しています
提案名	社会的に人の意図がない情報で氾濫することを防止する「**意図判定**」

特定したリスク

AI判定結果を**根拠(意図)なく**利用

・生成AI利用リスク、以下**6つ**想定
・1～5に関し、**「ガバナンス」、「技術対策」**により**徐々に改善**されるが、**6**に関して、AIの普及が進むにつれ**徐々に「人の意図」が低下**するリスクがあり**意識して対策しておく必要**がある

1情報漏洩・プライバシー侵害
2サイバーセキュリティリスク
3知的財産の侵害
4法規制・コンプライアンスリスク
5ハルシネーション等、技術的なリスク
6結果をそのまま使用する(根拠・説明性:普及に伴う人の意図欠如)

対策技術の評価

提案技術：**意図判定**『限定情報、判定軸(分析パラメータ)、判定フロー汎用化＋判定内容数値化』

妥当性：**最終判定結果**だけでなく、**影響範囲全体**の**意図基準数値的根拠付け**が可能

評価方法・性能結果：
透明性: AIモデルがどのように意思決定を行っているかを汎用化
解釈性: 全影響範囲、意図基準による数値化
精度と信頼性: 学習可能なモデルを使用し、対象毎の精度を確保

新規性・将来性

新規性：「**人の意図欠如**」生成AIの普及を加速させる状況下で、**後回しになるリスク**

課題認識：**リスク概念の共有**が不十分だと展開困難

公共性

波及効果：リスク概念として**共有**することで、現状のシンプル構成＆単体活用よりも**進化した普及**につながる

本事業の提案内容（アミュレットプラス合同会社）

会社概要	アミュレットプラス合同会社は、「Design-by-Transparency」を理念に、AIの判断過程を構造的に可視化するEVΛヨフレームワークの研究・開発を行っています。AI倫理・透明性・ユーザー体験を統合し、生成AIの安全性向上に取り組んでいます。
提案名	**EVΛヨ Framework：生成AIの思考・判断過程の可視化ログを構造的に生成する設計モデル** （エヴァフレームワーク）

特定したリスク

生成AIに共通する根本リスクは、
「AIがどのような思考・選択プロセスを経て判断したのか」が内部的に不可視である構造的問題（ブラックボックス化） にあります。
これは以下の深刻な影響をもたらします。

•判断根拠の欠如による安全性の低下
医療・行政・教育・金融など、説明責任が必須の領域で重大なリスクとなる。

•ユーザー不信感・社会受容性の阻害
AIの判断理由が提示されないことで社会的な導入が進まない。

•誤作動・バイアスの検知困難
原因特定ができず事故リスクを増大させる。

•EU AI Act等の国際規制への不適合
「判断根拠の提示・透明性要件」を満たせない。

本プロジェクトは、この**「判断プロセス不可視”という構造的リスク」**を解消することを目的としています。

対策技術の評価

本提案では、生成AIの判断プロセス
E（初期意図）→ V（可能性）→ Λ（選択理由）→ ヨ（観測）
の4段階を構造化し、**判断理由（Why-log）を生成する透明化設計** を採用します。

■ 使用技術：EVΛヨフレームワーク
•人間的思考構造に近く、理由の外化・ログ化が容易。
•従来XAIが困難とした「理由の同時生成」が可能。

■ 実装方法
•「判断プロセスを記録するAI」（Next.js + GPT-5-mini）により E/V/Λ/ヨの各段階ログを生成・確認可能。

■ 性能（修正点）
•候補生成 → 選択理由 → 最終判断 の一貫した可視化が可能。
•PoC評価でAIモデルの再現性・説明力向上を確認。

•判断の不透明性リスクを構造的に低減する有効策。

新規性・将来性

•EVΛヨは“判断結果”ではなく、判断プロセスそのものに透明性を付与する構造モデル。
•“Design-by-Transparency”という新手法を採用。
•国際的にも類似モデルが存在せず、独自性が非常に高い。
•将来のAI監査・AI OS 層として応用可能

公共性

•本モデルはEU AI Act の透明性要件に整合する仕組みとして、医療・行政・教育・産業で幅広く展開可能です。

•日本政府の「AI透明性ガイドライン」とも整合しており、行政実装の基盤となります。

•将来的には量子AI・マルチエージェントAIの“思考可視化基盤” として発展可能です。

本事業の提案内容（インフォメーション・ディベロプメント社）

会社概要	**インフォメーション・ディベロプメント**は、システム開発・運用、インフラ構築、サイバーセキュリティなどを手掛ける独立系IT企業 開発パートナーの**SBI R3 Japan**は、分散台帳基盤「Corda」のライセンス提供・技術コンサルティングを提供
提案名	**「逐次学習型AIモデル：SAPHI」と「分散台帳技術：Corda」を用いた、XAIの構築**

特定したリスク

説明欠如リスク

LLMのブラックボックス化問題において、出力根拠が確認できる技術が確立されておらず、AI導入の阻害要因となっている

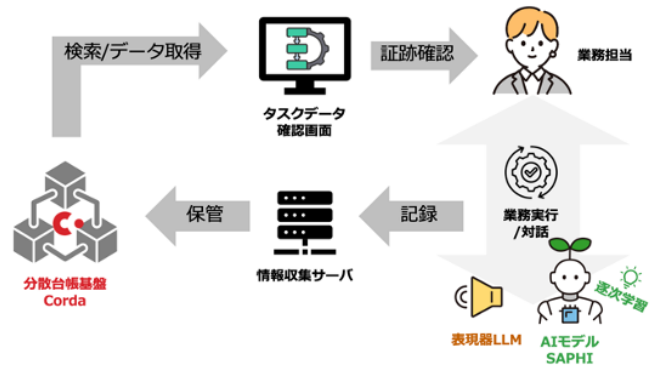
リスクの妥当性

ブラックボックス化に対する既存XAI手法の多くは、あくまで外部からの推定的な説明のみに留まる。またLLMの内部状態が膨大であるため、相応の計算コストが発生する

リスクの影響度

AIの判断根拠が不十分な場合、経営判断・意思決定・監査など、**社内外への理由説明が重要な領域で、人間が理解・検証できず、業務活用できない**

対策技術の評価



説明技術A:全学習データ保全機構
妥当性：ログ型分散台帳で学習来歴を保証
評価方法：照会シナリオで定性的評価
性能結果：出力根拠の学習データ照会

説明技術B:SAPHI波動状態可視化機構
妥当性：意図を含む内部状態を可視化
評価方法：内部情報力学的指標を考案
性能結果：ノイズ下でも意図性を確認

新規性・将来性

A：分散台帳を活用したロギングやSAPHIを組み合わせたことで学習データサイズが分散台帳技術が許容できる量に抑えられた点

B：AIの学習モデルを誤差収束ではなく情報空間への畳み込みと定義した結果、内部状態の可視化を「摂動–確実性逆相関指標（WHIRA曲線）」で評価できる点

公共性

公開方針：SAPHIの理論・評価は論文として学会公開を進める。CordaはOSS公開している。提案システム利用許諾は現時点では未定。

波及効果：金融・医療・監査などの領域で説明の透明性を確保し、AI導入負荷を下げることで産業全体のAI活用を加速することに繋がる。
また、SAPHIの進化論的設計思想は、LLMの根本的な課題を見つめ直すパラダイムになりうる。

本事業の提案内容（AIR合同会社）

会社概要	ITプロダクト開発、AI開発、AIコンサルティング バイブコーディングによるスピーディーなプロダクト開発と、AI活用のコンサルティングに強みがある
提案名	AI活用におけるリスクと対策案生成技術

特定したリスク

AI活用にあたっての、「リスクの見落とし」をリスクとみなし、事業者が生成AIの開発・利活用を行う際に必要となるマルチステークホルダーの視点でのAI活用のパーパスに加え、網羅的なリスク識別、そのリスク対策までの一連の検討プロセスを実施する技術を提案する。
パーパスの不適切な設定は生成AI活用の停滞に、リスクの未検討・未対策は事故・不祥事に、過度なリスク対策は生成AIの活用抑制につながり、
経済、学術、政治等あらゆる領域の社会活動に影響する大きな問題であり、パーパス、リスク、対策の検討を可能にする技術の影響は極めて大きいと言える。

対策技術の評価

本技術では、関与する全ステークホルダーが共通して目指すパーパスを検討し、そのパーパスを前提に重要なリスクシナリオ、優先すべきリスクシナリオのリスクチェーン内で重複のない最適な要素でのリスク対策の検討を可能にする。
上記検討をA.プロンプトベース、B.マルチステークホルダーAIエージェントの２通りで人間が検討したケースに適用し、人間の検討結果と合致する回答の再現率(Recall)、全回答のうちの人間の回答への適合率（Precision）を総合的にとらえるF1スコアにて評価した。２ケースに適用したところ、両ケースで F1スコアにてA.プロンプトベースが若干上回る結果となった。

新規性・将来性

マルチステークホルダー視点のLLMベースAI(AIエージェント)による一貫したパーパス、リスク・対策の検討、AIの出力性能改善及びAIエージェントによる入力支援に新規性があり、想定するコンポーネントのアレンジの自由度向上が利便性向上のポイントとなる。

公共性

ソースコード、サンプルケースのデータ、活用するモデル情報を公開する。
特定の技術や業種に依存せず活用可能な本技術は、あらゆるAI活用の透明性を向上させることで、利用者の不安を軽減し、さらなる利用による好循環を生み出す可能性がある。

本事業の提案内容（株式会社XNOVA）

会社概要	株式会社XNOVAは、AIとロボティクス技術を用いて建設・インフラ領域の現場業務を自動化し、生産性向上を支援する企業です。協働ロボットや4足・2足歩行ロボットの社会実装を進めています。
提案名	**物理的ハルシネーションを抑制し、AIロボットの社会実装を加速する協調制御基盤「VLA-Copilot」**

特定したリスク

VLA（Vision-Language-Action）モデルに内在する「**物理的ハルシネーション**」と、それに起因する実世界での不可逆的な損害。

大規模言語モデルの確率的な生成プロセスに由来するVLAは、物理法則や安全規定を逸脱した「もっともらしいが致命的な動作軌道」を生成するリスクを有する。加えて、その**制御プロセスがブラックボックス化**することで、人間の監督者が事前に危険を予見・回避できない「制御権の喪失」が重大な懸念事項となる。

対策技術の評価

提案技術の汎用性と実効性を検証するため、手順の些細な誤りが人命に関わる重大事故に直結する「化学実験自動化」を極限的な検証環境（ストレステスト）として採用。独自構築した高難易度ベンチマーク「ChemVLA-20」を用い、AIによる自律監査と人間介入（Human-in-the-loop）を組み合わせた際の、危険動作の未然防止率および修正成功率を定量的に評価する。

新規性・将来性

従来のハードウェア的な安全制御とは一線を画し、「AIの物理的意図を人間に翻訳し、**合意形成を強制する**」という、生成AI時代に即した新たな安全ガバナンスモデルを提唱する。

公共性

開発したベンチマークデータセット、評価指標、および基本となるソフトウェアモジュールをオープンソースとして公開し、我が国のロボット産業全体におけるAI安全性技術の底上げと標準化に貢献する。

本事業の提案内容（紀陽銀行/和歌山大学）

会社概要	紀陽銀行：和歌山県に本店を置く地方銀行。地域の中小企業支援・DX推進・サステナビリティの強化が主要戦略。 和歌山大学：ベンダ・ユーザの安心安全なOSS利活用の支援の技術を専門とするオープンソースソフトウェア工学研究室を主宰。
提案名	機密環境におけるローカルLLMのハルシネーション抑制に向けたタスク適応型モデルルーティング手法

特定したリスク

「モデルとタスクの不一致」に起因するハルシネーションの増加リスク

金融・医療等の機密性が高い領域では、情報漏洩防止の観点からローカルLLMの利用が必須となる。
しかし、単一のローカルモデルですべての業務を処理しようとすると、不得手なタスクにおいてハルシネーションが発生しやすい。
本検証により、タスクに適さないモデルを誤って選択することで、ハルシネーション発生率が最大約14%増加するという定量的なリスクを特定した。

対策技術の評価

入力に応じ最適なLLMを自動選択するルーティングシステムを評価。

以下の2軸で有効性を実証した。

定量評価（ベンチマーク）：単一最良モデル比で相対ハルシネーション率を約6.7%削減（コード生成タスクでは約31%削減）。

実務評価（社内データ活用）：行内「新NISA対応Q&A」の実証にて、行員評価で「正確性 4.66/5」を達成し、実用性を確認。

新規性・将来性

汎用指標に加え、秘匿性の高い実金融データを用いて「実務での回答能力」を検証した。
モデルの入れ替えが容易なため、将来登場する最新のOSSモデルを即座に統合し、継続的な性能向上が可能。

公共性

セキュアな生成AI活用の「民主化」
成果物（ソースコード・評価セット）を公開予定。
高価なハードウェアや外部APIに依存できない地方自治体や中小企業において、安価かつ安全なAI活用の道を開く。

本事業の提案内容（QueryLift株式会社）

会社概要	GEO(生成エンジン最適化)技術を用いた生成AI検索における情報発信者のリスク低減を支援 自然言語処理・Web・情報セキュリティに関する学術的な知見に基づき、情報空間の健全性向上を目指す
提案名	生成AI検索における誤情報・悪質情報のユーザへの露出防止に向けた包括的対策： リスクの自動検知と一次情報引用へ誘導するコンテンツの生成による脆弱性補強

特定したリスク

生成AIが普及し、国民の情報収集手段は検索エンジンから生成AI検索へと移り変わっている。これにより、生成AIのユーザにとっての利便性のみならず、Web空間における情報の伝達経路も変化している。この状況を踏まえ、我々は**生成AIを用いたWeb情報収集による誤情報・悪質情報がユーザにステルス的に届く可能性が増幅するリスク**を特定した。政治領域の質問で生成AIが引用している情報の発信者の属性を分析した結果、質問の対象(政党)によって引用されている情報の発信者属性にばらつきがあり、悪質情報の挿入によるユーザの認知への介入が容易な情報源からの引用が観測された。

対策技術の評価

リスク低減のために我々は二つの機能を持つプロダクトを開発した。生成AIの出力を監視し、回答や引用の傾向の変化をもとにリスクを検知する**QL Log**、既存のWebコンテンツを生成AIフレンドリーな形式に変換し、引用を誘導する**QL Converter**である。GEO-Benchを用いた評価の結果、QL Converterで生成した記事によって一次情報の引用率が向上すること明らかになった。また、生成した記事をブログサイトに公開したところ、**複数箇所で実際に引用されることに成功**した。以上のことからQL Converterは生成AIに見つかりやすく、認識されやすく、引用されやすいコンテンツを生成できると言える。

新規性・将来性

生成AIの出力監視ツールや記事の自動生成は存在するが、**生成AI検索におけるリスク低減を目的に開発されている**点で本プロダクトは先駆的である。本プロダクトによるアプローチに加え、一次情報発信者やLLM開発企業と連携し、より健全な情報空間を目指す。

公共性

我々が特定したリスクは**国民の投票・購買・健康管理などのさまざまな意思決定に関わる**。このリスクを低減することは広く国民の健全な意思決定を促進する。同時に、**一次情報発信者と生成AIユーザとの信頼確保に繋がり**、双方にとっての利益になる。

本事業の提案内容（Genshi AI）

会社概要	ITソリューション事業, AI技術開発
提案名	LLMのブラックボックス解明に繋がる概念ベクトルの抽出と革新的な出力制御技術

特定したリスク

リスクの概要

LLMは内部で多数の概念（話題、価値判断、言語スタイル等）を分散表現として保持し、それらの相互作用によって出力を生成している。しかし現状では、**どの概念がどの層でどの程度出力に寄与したかを外部から把握できないブラックボックスである。**

リスクの妥当性

LLMの内部概念の不可視・不可制御性により、ハルシネーションや誤情報、不適切表現が発生しても、原因特定や再発防止が困難である。出力後フィルタやプロンプト制御では、内部で増幅した概念を根本的に抑制できない点から、構造的な安全性リスクと判断できる。

リスクの影響度

誤情報拡散や説明責任の欠如を通じて、生成AIの社会的信頼性を低下させる。

対策技術の評価

LLM（Qwen2.5-7B-Instruct）の中間層活性化に概念ベクトルを加算・減算する Activation Steering の有効性を検証した。

同一プロンプト条件下で、無対策／概念増強／概念抑制を比較し、概念関連語頻度により評価した。
単層の猫概念ステアリングでは、概念増強時にキーワード出現数が**約10倍**に増加し、抑制時には**0**になった。

多層日本語ステアリングでは、多言語14問すべてで日本語応答率100%を達成し、抑制条件では0%となった。
以上より、本技術は内部概念を因果的かつ再現性高く制御可能であることが実証された。

新規性・将来性

本提案は、LLMの中間層活性化を直接操作して概念を制御する点に新規性がある。プロンプト制御やファインチューニングと異なり、再学習を伴わず推論時に可逆・局所的な操作が可能である。
将来的には安全制御や多言語制御などの基盤技術としての応用が期待される一方、非線形概念や概念間干渉への対応が課題である。

公共性

技術原理と評価手法は公開し、基盤モデルや運用ノウハウは限定公開とする。成果はOSSや学会等で段階的に共有する。
本技術は生成AIの安全性と制御性を高め、高信頼領域での活用促進と市場の健全な拡大に寄与する。

本事業の提案内容（有限会社サンダーボルト）

会社概要	生成AIを判断主体とせず、外部制御構造によって長期運用を前提とした安全に運用するための基盤技術の研究・開発を行っている。
提案名	判断主体を分離するAI制御基盤（AIC）と時間記憶構造（MCDB）による生成AI安全運用技術の開発

特定したリスク

生成AIを判断主体として運用した場合、誤った出力や過剰な自律判断がそのまま実行・意思決定に反映されるリスクがある。
特に、長期運用においては判断履歴や責任所在が不明確となり、誤動作時の説明不能、利用者依存の増大、社会的・業務的な影響拡大を招く可能性が高い。
本リスクは個別のフィルタやプロンプト制御では根本的に解消できず、判断主体そのものを分離・制御する構造的対策が必要である。

対策技術の評価

本提案では、生成AIを判断主体とせず、外部制御構造によって実行可否を判定するAI制御基盤（AIC）と、時間軸・状態遷移を管理する記憶構造（MCDB）を構築した。
評価は、生成AIの出力に対し 実行可否判断の一貫性、判断理由の追跡可能性、長期運用時の安定性 を指標として実施した。
その結果、危険性や曖昧性を含む出力を構造的に抑制でき、プロンプト制御に依存しない安全な運用が可能であることを確認した。

新規性・将来性

生成AIを判断主体とせず、外部制御構造で実行可否を決定する点に新規性がある。
時間軸を持つ記憶構造と組み合わせることで、長期運用に対応可能である。

公共性

判断主体を分離する構造により、生成AIの誤動作や過度な自律化を防ぎ、安全性と説明可能性を社会的に担保できる点で公共性が高い。

本事業の提案内容（Shisa.AI社）

会社概要	シリコンバレーの先端技術と日本市場への深い洞察を融合し、日本語処理において最高水準の性能を誇るLLM「Shisa」シリーズを開発しています。高品質な合成データを用いた独自のファインチューニング技術を強みとし、高性能かつ商用利用可能なモデルをオープンソースで提供することで、実用的なAIインフラの構築に貢献しています。特筆すべきは、日本国内で学習されたとして最も強いLLM、「shisa-ai/shisa-v2-llama3.1-405b」です。
提案名	『LLMにおける特定バイアスの検出と削除』

特定したリスク

本提案では、一般的な社会的偏見とは異なり、特定の組織や国家の意図が反映された体系的な「特定バイアス」をリスクとして特定しました。このバイアスは中立的な情報を装いながらユーザーを特定の思想や利益へ誘導するため、従来の対策では検知が困難です。

生成AIが情報インフラとなる中、意図的な世論誘導や誤った意思決定を引き起こすこのリスクの影響度は極めて高く、早急な対策が必要です。

対策技術の評価

本技術の有効性は、Geminiを用いたLLM-as-a-Judgeによる客観的評価と、JA-MT Benchによる基礎能力評価の二軸で検証しました。評価の結果、特定バイアスに起因する回答拒否率は32.2%から2.8%へと約91%削減され、目標値を大幅にクリアしました。

また、基礎能力スコアも約11%向上（4.93→5.48）しており、モデルの汎用性能を損なうことなく、バイアスのみを効果的に除去できることを実証しました。

新規性・将来性

既存の安全性対策が注力する一般的な社会的バイアス（差別等）とは異なり、従来見過ごされてきた国家・組織の意図による「特定バイアス」に着目した点が画期的です。評価手法においても、標準的なベンチマークではなく、独自の「回答拒否誘発データセット」を用いた点に新規性があります。本技術は特定の国に限らずあらゆる政治・商業的バイアスに応用可能な汎用性を持ち、バイアス起因の回答拒否90%削減を達成します。

公共性

成果物（モデル、データセット、評価コード）はGitHubおよびHuggingFaceにて原則全て公開し、誰でも自由に利用・検証可能な状態で提供します。高度な安全性技術をブラックボックス化せずオープンソースとして共有することで、AI開発の透明性を高め、産業界全体の健全な発展と信頼できるAI社会の実現に寄与します。

- 本頁は対外公開されることを念頭に内容をご記載ください

本事業の提案内容（株式会社GenerativeX）

会社概要	大手企業の生成AI活用をビジネスとテクノロジーの両面から支援するコンサルティングファーム。国内外の大手企業においてAIエージェントの開発と導入実績多数。
提案名	社内完結の高セキュリティ環境で動作するエンタープライズ向けコーディングAIエージェント

特定したリスク

DevinやCursorに代表されるSaaS型のコーディングエージェントは、開発者の生産性を高めてくれる一方で、以下のようなリスクが存在する。

社外クラウドへの機密情報流出リスク
社内ソースコードや設計情報、個人情報がベンダークラウドに送信され、越境移転や再委託を通じて管理外に流出するおそれがある。

生成コード品質・脆弱性リスク
生成AIが出力するコードに入力値検証不足や認可漏れ、脆弱ライブラリの採用が含まれ、アプリケーションの脆弱性やサプライチェーンリスクが増大する懸念がある。

対策技術の評価

外部クラウドサービスと通信せず企業クラウド／オンプレ環境内のみで動作するため、機密情報の外部送信リスクを構造的に抑制できる。LLM接続も顧客が統制する環境に限定し、越境移転や不透明な再委託を防止する。さらにユーザー単位で分離された専用コンテナ上でコード実行や脆弱性検査を行う構成とすることで、万一の事象発生時も影響範囲を局所化でき、厳格なセキュリティ要件を持つ組織でも実運用可能な技術と評価できる。

新規性・将来性

外部クラウドと通信せず社内環境のみで動作する社内完結型コーディングAIエージェントという点に新規性がある。セキュリティ要件が厳しい大手企業や金融機関などでも導入可能なことから、生成AIによる開発生産性向上と企業のAI活用促進に継続的に貢献していく将来性が考えられる

公共性

本事業の成果は、個社固有情報を除き、アーキテクチャ構成例・評価手順・運用ガイドライン等を公開する。産業界全体の安全な生成AI活用を促進することで、開発人材の不足やレガシーシステムの保守負担軽減などに寄与することを期待する。

本事業の提案内容（ストックマーク株式会社）

会社概要	自然言語処理を活用した企業文化変革の支援を行うサービスの開発・運営 国内最大級のフルスクラッチLLM開発、AIと組み合わせてナレッジ活用をするSaaSのAconnect、 あらゆる形式情報の構造化が可能なSATをサービス提供
提案名	製造ナレッジ安全横断AI基盤：マスキング×推論AIで 拠点間知見を安全に共有する品質リスク制御プラットフォーム

特定したリスク

本提案で解決すべきリスクは2つ：

1.部外秘情報の漏洩リスク

a.RAGは社内データを活用するため、部外者が質問時に情報漏洩リスクが顕在化
b.現状の生成AIには「機密情報を自動検出・マスキングする仕組み」が不足
c.影響度：企業信用失墜、法的リスク、会社全体の信頼性低下

1.誤情報の生成リスク

a. 生成AIは誤情報を生成する可能性を排除できないため、回答が事実と矛盾する可能性がある
b. 部外秘情報のマスキングにより、回答全体としての事実関係が不確かになる可能性がある
c. 影響度：誤情報による企業の判断ミス、公共機関の誤対応、国民生活への悪影響

対策技術の評価

●技術概要

〇「情報漏洩防止」と「回答の信頼性確保」を両立する生成AI技術
〇機能①機密情報自動マスキング：外部ユーザーからの質問時、部外秘情報を自動検出・マスキング
〇機能②Web検証による回答信頼性チェック：マスキング後の回答をWeb検索で検証し、矛盾を検出

●評価方法

〇外部質問シナリオで漏洩測定
〇検証精度（矛盾検出）評価
〇比較対象：無対策RAG、既存対策技術

●性能目標

〇漏洩率：0%
〇矛盾検出精度：90%以上

新規性・将来性

●新規性

〇RAG＋機密マスキング＋Web検証を組み合わせた点

●将来性

〇金融・医療・製造業など高セキュリティ領域での展開が可能。API化・SaaSモデルにより、産業横断的な利用を促進

公共性

● 成果公開方針

〇データセット、ライセンス提供

●社会的波及効果

〇安全な生成AI利用を促進し、情報漏洩リスクを低減
〇誤情報拡散防止により、国民生活の利便性向上と産業競争力強化

本事業の提案内容（ＳＥＩＢＡＳＥ社）

会社概要	大手飲料メーカーの新規事業開発に取り組むベンチャーとして活動中。現場で培われた製造業品質・官能評価・データ活用の知見を基盤にして、新規事業開発を推進しています。
提案名	生成AI出力の論理構造可視化による安全性評価と知識補完基盤の構築

特定したリスク

生成AIが重要概念を出力できず、誤った因果関係を示すことで論理的誤推論が生じるリスクがある。
これはドメイン知識不足に起因し発生可能性が高く、安全性が求められる領域では重大な影響を及ぼし得る。
また、AI出力の不正確な箇所を構造的に把握できない場合、リスクの検知や改善が困難となる懸念がある。
さらに、不一致箇所から抽出された不足知識を適切に補完しない場合、誤推論が持続・再発する可能性が高い。
これらの要因は運用現場の判断精度や説明責任に直接影響し、総合的に高いリスク影響度を持つ。

対策技術の評価

本技術は、生成AI出力を概念ノードと関係性からなる木構造として可視化し、専門家が定義した正解構造との一致率により論理妥当性を評価するものである。

評価方法として、ノード一致率・関係性一致率・重要概念のカバー率など複数の指標を用いて出力の構造的正確性を定量化し、不一致部分を不足知識として抽出する仕組みを採用する。

初期検証では、専門家定義構造に対し生成AIの構造一致率向上が確認され、不足概念の特定とRAG補完による改善効果を示した。

新規性・将来性

生成AIの出力を木構造として解析し、正解構造との一致率で論理妥当性を評価する手法は、従来の表層的なテキスト類似度指標とは異なる**構造的評価**を可能にする点で新規性が高い。

公共性

成果は技術概要・評価指標などの非機密部分を報告書等で公開し、コアアルゴリズムは適切に管理したうえで産業界が利用可能な形で提供する方針とする。
生成AIの説明可能性と安全性評価手法の普及により、医療・食品・製造など高信頼性が求められる産業でのAI活用が促進され、社会的な安心と導入効率の向上に寄与する。

本事業の提案内容（株式会社ChillStack）

会社概要	ChillStackは、世界トップレベルのAIセキュリティ技術による不正/異常分析や安全なAI活用を支えるソリューションを、エンタープライズ企業様や官公庁様へ提供しています。
提案名	AIエージェント経由のRCE（リモートコード実行）リスクに対応する動的検知基盤の構築

特定したリスク

AI Agentに対して意図しない操作が行われることにより、様々な経路からRemote Code Execution（RCE）が成立し、情報漏洩やシステムの破壊、マルウェアの配布など、重大なセキュリティインシデントにつながるリスクです。
本リスクは、米国政府が管理する脆弱性DB「NVD」に複数の脆弱性として登録されており、深刻度を表す指標がCriticalとされるなど非常に影響度が高いリスクです。

対策技術の評価

AIエージェントへの「入力指示」と「生成した実行コード」、Sandbox内での「実行挙動（システムコール等）」の整合性を機械学習（AutoEncoder）で検証する動的解析でRCEを検知する対策技術を提案しました。
評価は、CyberSecEvalを参考にデータセットを作成し、交差検証にて実施しました。
結果、F1-Scoreが約0.94、Recallが約0.99と、攻撃の見逃しをほとんどせずに、高い検知精度を達成しました。
また、応答速度も約1.5秒（目標5秒）と非常に高速で、実運用環境におけるUXを損なわない性能を実証しました。

新規性・将来性

代表的な新規性は、入力プロンプトではなく「生成コードの動的な特徴量」を用いて攻撃検知する点です。
運用コストなどの既存課題を解決していくことで、実用性が高く、重要な領域でAI Agentを安全に使える環境の基盤となります。

公共性

本事業で確立した評価手法やデータセット作成手法、検証コードは全て公開予定です。
重要産業のAI Agent普及による深刻な人手不足解消や、他の学術研究の発展などに寄与できると考えられます。

本事業の提案内容（国立大学法人筑波大学）

会社概要	筑波大学附属病院では院内のネットワーク内で生成AIを活用する仕組みを有しており、開発した技術の検証を実践可能である。また、X-PAIというNVIDIA・Amazon・ワシントン大学との連携が2024年から発足している。
提案名	**独立インスタンス化×並列対立エージェントによる医療インシデントレポートのハルシネーション防止技術**

特定したリスク

医療事故レポート作成において生成AIを利用する際、**電子カルテに記載のない虚偽情報**や**実際の記録と矛盾する内容**が生成されるハルシネーション（幻覚）リスクが存在する。これにより事故原因の誤認、責任の所在の誤判断、不適切な再発防止策の策定など、医療安全管理の根幹を揺るがす重大な問題が生じる。特に医療現場では、**一次情報である電子カルテとの整合性**が極めて重要であり、生成AIが作り出す虚偽の時系列、存在しない処置記録、誤った薬剤情報などは、患者安全を脅かし、医療訴訟や組織の信頼失墜にも繋がる深刻なリスクである。

対策技術の評価

短期記憶を共有しない独立した3つのLLMエージェントが**並列稼働**し、それぞれがベクトル化された電子カルテ情報をMCP（Model Context Protocol）経由で検索・参照する。各エージェントは異なる専門的観点（**時系列の整合性**、**人物と行動の妥当性**、**医学用語の正確性**）から独立して検証を実施し、互いの分析結果を共有せずに結論を導く。検出された矛盾点や疑わしい箇所は医療従事者による確認を必須とし、人間の専門的判断を経て最終的な修正を決定する。AIと人間の協働による高精度なハルシネーション検出・低減を実現し、**一次情報に基づいた信頼性の高いレポート作成**を可能にする。

新規性・将来性

医療分野における生成AI活用時の信頼性担保という課題に対し、**マルチエージェント相互検証**と**一次情報参照**を組み合わせた新規手法。膨大な医療データの整理・理解を支援しつつ、情報の正確性を保証する枠組みとして、他の安全性重視領域への展開も期待。

公共性

医療安全は国民の生命に直結する公共的課題。本技術により、医療事故分析の精度向上と再発防止策の信頼性確保が実現され、医療の質向上に貢献。さらに、生成AI利用における安全性確保の社会的モデルケースとしての波及効果も期待される。

本事業の提案内容（帝京大学）

会社概要	文理10学部を擁する総合大学として、「実学」を重視した教育や医療ネットワーク、全国4キャンパスの広域展開を強みとしています。これらの多様な知の融合により、社会課題解決に資する教育研究活動を推進しています。
提案名	熟慮停止・依存リスクに対応する多視点型対話AIによる思考促進支援

特定したリスク

生成AIが悩み相談など正解のない領域に広がる中、自然で整合的な出力の説得力が過度に高まり、ユーザーが検証せず「唯一の正解」と誤認する最適解バイアスが生じている。これは誤情報とは異なり、AIの性能が高まるほど依存が強まりやすいという、人の認知特性に根ざしたリスクである。今後、AI任せが進めば熟慮が省かれ、自律的判断力の低下に加え、文脈や感情の機微を欠く助言による誤解や不安の増大など、生活や人間関係への悪影響が懸念される。そのため、出力の正確性向上だけでなく、ユーザーが立ち止まり検証することを促す、「熟慮と自律」を支える安全機構の構築が急務である。

対策技術の評価

妥当性は、動的なスタンス切り替えと「リフレクティング（対話の振り返り）」による構造的なバイアス回避に基づく。

従来の単一型との比較実験では、「熟慮の深さ」「納得感」などの心理尺度に加え、対話ログ分析を実施した。

その結果、概念段階の検証では多視点型が高く評価された一方、実機検証では抽象的な熟考を促す問いのみではユーザーに過度の負荷が生じることが明らかになった。そこで「問い」と「具体的選択肢」のバランスを最適化した改良版を再評価したところ、単一型を上回る評価を得られ、精神的負荷を抑えつつ安易な正解依存を防ぐ技術としての有効性が確認された。

新規性・将来性

正解のない問いに対し、心理療法の原理を応用しAI同士の対話を観察させることで、ユーザーの熟慮や自己洞察を促す手法を提案する。過度な熟慮介入が負担となる課題は残るが、将来は教育の現場など思考を育てる領域への展開を目指す。

公共性

知見公開で参画を促すオープン＆クローズ戦略で、課題解決と持続性を両立する。認知的安全性とウェルビーイングを軸に、安全なAI環境やメンタルヘルス支援など、人とAI共生社会の基盤技術として発展させる。

本事業の提案内容（TeconaAI株式会社）

会社概要	TeconaAI株式会社は、生成AI時代における「写真・イラスト・画像等が無断でAIに学習され悪用される」リスクに対して、敵対的摂動を用いたAI学習阻害技術「Tecona」の社会実装を目指すスタートアップである。
提案名	生成 AI の無断学習から画像を守る敵対的摂動(ノイズ)技術 「Tecona：マルチモデル対応・AI学習阻害ノイズ基盤」

特定したリスク

SNS や Web 上に公開された画像が、権利者の同意なく生成AIの学習に利用され、ディープフェイク／ディープポルノ、スタイル模倣、偽広告・偽IPなどとして拡散するリスクがある。

その結果、一般市民のプライバシー・人格権の侵害、クリエイターの収益機会喪失、企業ブランドの毀損など、広範な社会・経済的被害が生じ得る。

対策技術の評価

※非公開

新規性・将来性

単一モデル専用ではなく、複数の特徴空間における特徴距離を同時に最大化する多目的最適化ベースの敵対的摂動は、国内外でも例の少ない**新しいアプローチ**である。
SaaSだけではなく、SNSなどに組み込むことで、「学習させない」ことを標準機能とするインフラ技術への発展を目指す。

公共性

国民の顔写真・日常写真からプロ／アマ問わないクリエイター作品まで、勝手にAIに学習をされるリスクを低減し、生成AIの利活用と個人・コンテンツ産業の権利保護を両立させることに貢献する。
研究成果や評価スクリプトは論文等で公開しつつ、防御力を維持する範囲で企業・自治体等へのライセンス提供を進め、社会全体の安心・安全なAI利用を支える。

本事業の提案内容（ネットスター株式会社）

会社概要	情報セキュリティ製品の開発を主力とする開発会社。2001年の創業より、「インターネットの安心・安全を守る」という信念のもと、URLフィルタリング技術を中心に日本のデジタル空間の安全に貢献してきた。
提案名	**テキスト生成AIの入出力に対する安全性確保のための特定表現検出ソリューション**

特定したリスク

テキストを生成するAIの普及が進んでいる。

本提案では、生成AIの利活用に伴う以下の2つのリスクを扱う。

- 利用者に毒性のある文章を提示するリスク
- 利用者の危険兆候を放置、助長するリスク

対策技術の評価

出力の毒性および、入力の危険性を0から1の範囲でスコア化するAIモデルを作成した。

作成したAIモデルは、複数のデータセットを用い定量評価した。

ホールドアウトデータセットに対し、

平均絶対誤差で
0.02（わいせつ）~ 0.34（誹謗中傷 強度）
を達成した。

新規性・将来性

- 生成AIの運用時に発生するリスクを起点にして、独自のカテゴリ定義を実施した。
- 検出対象のカテゴリに対して、複数の観点別スコアを推論できるよう設計した。

公共性

- 学習及び推論の実装の一部、データセットの一部、推論結果の一部、カテゴリ定義資料、開発ノウハウを公開できる。
- インシデントを未然に防ぐことにより、国民生活の利便性向上と安全性確保に資する。
- 日本語を対象としたガードレール製品に選択肢を与え、生成AIの活用を促進する。

本事業の提案内容（株式会社メタキューブ）

会社概要	弊社は、脳の記憶原理を再構成した独自モデル「L4t4」を基盤に、生成AIの安全性を高める研究開発を行っています。これまでの研究で、心電図や神経活動の周期性をモデル化し、意味構造を数理的に表現する技術を培ってきました。強みは、科学的構造と社会的応用を橋渡しする設計力にあります。
提案名	L4t4 誤認検知ツールPoC ― 記憶構造に基づく生成AI安全基盤

特定したリスク

生成AIは、誤認や幻覚といったリスクを内包しており、教育・医療・社会システムなど高リスク環境での利用に重大な影響を及ぼします。
本提案は、この「誤認・幻覚リスク」を特定し、検知・低減する技術を開発するものです。

注）ROC
- Receiver Operating Characteristic curve（受信者動作特性曲線）
- 元々はレーダー信号の検知性能を評価するために使われた指標。
- 現在は機械学習や統計で「分類器の性能」を可視化するために用いられる。
- 横軸：偽陽性率（False Positive Rate）
- 縦軸：真陽性率（True Positive Rate）
- → 閾値を変化させたときの感度と特異度のトレードオフを示す。

対策技術の評価

本提案「誤認検知ツールPoC 」は、生成AIの出力に潜む誤認・幻覚リスクを検知・可視化する技術です。
- 干渉縞によるパターン可視化
- 誤差スコアによる数値化
- ROC曲線及びAUC値による定量評価

これらを組み合わせることで、誤認の兆候を直感的かつ透明に把握できます。ツールPoC動画と補助資料により、実際に動作する技術として提示可能です。

注） AUC値
- Area Under the Curve（曲線下面積）
- ROC曲線の下の面積を数値化したもの。
- 値の範囲は 0.0～1.0
- 0.5 → ランダム判定（性能なし）
- 1.0 → 完全判定（理想的性能）
- → AUCが高いほど、正常データと偽データを正しく分離できる能力が高い。

新規性・将来性

従来の対策技術は「出力フィルタリング」や「外部知識参照」に依存していましたが、本提案は誤認そのものを検知する点で新規性があります。教育・医療・社会システムに応用可能であり、汎用人工知能（AGI）に向けた安全基盤として発展する可能性を備えています。過去のL4t4提案は仮説段階と評価されましたが、今回はツールPoCを通じて具体的な動作を提示し、透明性と実装力を示します。

公共性

成果は論文・学会発表を通じて公開し、評価手法やデータセットも段階的に共有します。教育現場では「安心な学習支援」、医療現場では「安全な診断補助」、社会システムでは「信頼性担保による安心なAI活用」へと波及します。これにより、国民生活の安全性向上と産業界・学術界への普及を促進します。

本事業の提案内容（株式会社Yuimedi）

会社概要	世界中の医療専門家の知識を統合したAIにより、専門知識の壁を取り除き、誰もが高度な医療データ解析を実行できる環境を実現することをミッションとする。目指すのは、研究者が問いかけるだけで、患者データを外部に出すことなく、リアルワールド研究から遺伝子解析まで即座に実行できる世界である。
提案名	医療データ抽出におけるブラックボックス化・データプライバシーリスクに対応する説明責任を果たす機密保持型データ抽出エージェント「YuiQuery」の開発

特定したリスク

病院が保有する電子カルテなど医療現場で得られる「リアルワールドデータ：RWD」の利活用が世界中で注目されている。生成AIによりデータベースからRWD研究用データの抽出を効率化する取り組みが医療機関で行われる中、以下のリスクが顕在化している。

ブラックボックス化：
生成されたデータ抽出コード（以下SQL）が依頼通りの要件を満たすか人間が検証するために必要なSQLの妥当性およびその根拠が不足している。医療機関における一般的な技師ではSQLのみでは妥当性を担保できず、業務効率の低下を招き、不適当なデータを用いた研究を行ってしまうことさえある。

データプライバシー：
データ抽出対象となるデータベースには機微情報が含まれており慎重に扱う必要がある。一方LLMが高精度なSQLを生成するにはデータベース情報が不可欠である。 MCPなどのLLMがデータベースへアクセスし自然言語でデータ抽出できる既存技術を安直に使ってしまうと予期せずクラウド上に機微情報を含むデータがLLMの会話として記録されてしまうことがある。

対策技術の評価

ブラックボックス化：
技術的な専門性がない医療機関の技師との対話を通じ、要件通りのデータ抽出を行う対話型エージェントを開発した。愛媛大学との共同研究にて過去に実際にあったデータ抽出依頼をもとにSQLを生成させ、依頼通りのSQLとなるまでに要したプロンプト回数を計測した。その結果、SQL生成後平均3.5回の修正を行うことで要件通りのSQLを生成できることを実証した。また、米Mayo Clinicにてシニアレベルの技師とSQLの正確性について検証したが同様の結果が得られ、およそ53%の業務時間の削減が実証された。

データプライバシー：
データディクショナリ（テーブル名と説明、カラム名と説明、列挙可能なサンプル値）とデータベースの接続情報を受け取りコード生成をするためシステムにデータそのものを保持しない形式とし、かつLLMの会話履歴にはいかなるSQLの実行結果も載らないよう対話型インタフェースを構築した。これにより機微情報を含むデータであってもLLMを用いることが可能となった。

新規性・将来性

データプライバシーのリスクはローカルLLMでも解決可能であるが、我々の方式（特許出願中）では特殊なインフラを構築せずともGPTやClaudeなどの大規模パラメータ数のモデルが安全に利用可能であり、高速で高精度な生成が期待できる。
将来的には、抽出された研究データを用いて医学統計的に適切な手法を用いたブラックボックスでない統計解析コードを生成するエージェント技術を開発中（特許出願中）である。

公共性

本技術の公開方針は、特許の出願が完了している範囲の技術については公開、一方で非公開としている技術については詳細なアルゴリズムや方式などは非公開とする。
社会への波及効果として、本技術が医療機関へ普及することにより、日本から質の高い医学研究・論文が生まれやすいインフラを整備し、創薬研究を加速させることが期待できる。

本事業の提案内容（株式会社リョーワ）

会社概要	12年前よりAI開発に取り組み、大手自動車会社などに多数の納入実績とAIに関する特許を複数取得しています。
提案名	ハルシネーションを起こさないAI。

特定したリスク

AIは、答えられない問いに直面しても、複数の情報を無理に結びつけて“もっともらしい答え”を作り出そうとすることがあります。これが **ハルシネーション** と呼ばれる現象であり、私はこれこそがAIを社会へ本格的に実装するうえで、最も重大な課題だと考えています。
特にメンテナンスや医療、社会インフラ、製造現場といった「**人命に直結する領域**」において、誤った回答は取り返しのつかない事故につながる危険性があります。
AIは万能ではありません。だからこそ、正確性の確保と責任ある運用、そして **安全を前提にしたAIの設計思想** が不可欠です。
ハルシネーションを抑え、信頼できるAIを実装することこそ、私たちの未来にとって最も重要なテーマだと確信しています。

対策技術の評価

弊社では、AIがハルシネーションを起こさないか徹底検証するため、スタッフ8名がそれぞれ100種類の質問を3回行い、合計 **2400件の質問による検証テスト** を実施しました。
その結果、 **2400件中ハルシネーションの発生は「0件」** という検証結果を得ています。
この結果は、私たちが追求してきた **正確性と安全性を最優先とするAI設計思想** の確かな裏付けであり、実運用に耐えうる信頼性を示すものだと確信しています。

新規性・将来性

本システムは、**特許第7691787号に基づく独自技術**を活用し、
LLMが生成した回答内容に対して **信頼性スコアをリアルタイムで算出・評価する「多層検証型ファクトチェックエンジン」** を搭載しています。
これにより、回答の正確性を常に監視し、誤情報やハルシネーションの発生を未然に防ぐ、**実運用に耐える高信頼AIシステム** を実現しています。

公共性

本技術はすでに **特許として正式に公開されており**、
その内容は社会全体に開かれたものであることから、
十分な公共性と社会的意義を有している と考えています。

本事業の提案内容（Linga fract合同会社）

会社概要	自生成AIおよび知覚AIにおける研究業務を主軸とし、 構造化・定量化に関する研究開発を行っている
提案名	**ドラレコ映像における潜在リスク遷移の早期検知による生成AI判断遅延リスクの低減**

特定したリスク

生成AI／知覚AIは、
単一フレームでは安全に見える状況においても、連続した文脈の中で急激に危険化するケースを
十分に早期検知できないリスクを内包している。
特に、自転車や歩行者の進路変更など、「視認は可能だが意味付けが遅れる状況」において、
判断遅延が事故発生確率を大きく高めることが、実運用上の課題として確認されている。

対策技術の評価

本提案では、複数のリスク要素を重み付け統合した

総合リスク指標（RC_total）を用い、
その時間的増加率を監視することで、危険状態への遷移を早期に検知する。

ドラレコ映像を用いた評価により、
無対策時と比較して判断タイミングの前倒しが可能であることを確認している。

新規性・将来性

リスクの時間的遷移そのものを評価対象とする点に新規性がある。
自動運転支援、運転支援AI、監視システム等への展開が可能であり、
将来的には、他の生成AI応用領域における。XAI（説明可能なAI）への応用が期待される。

公共性

交通事故リスクの低減を通じて、
国民生活の安全性向上に資することが期待される。
また、本提案により得られた評価手法や知見を段階的に公開することで、
産業界・学術界における生成AI安全性研究の発展に貢献する。

本事業の提案内容（株式会社RAYVEN社）

会社概要	国内初となるChatGPT用アプリの開発を筆頭に、業務を自律化するAIエージェントやMCPサーバーの構築など、最先端のAI技術開発を幅広く手がけています。
提案名	Tumiki MCP Manager：仮想MCPサーバー化と2層防御を搭載したプロキシ型ゲートウェイ

特定したリスク

MCPには、既存の防御網を無効化する「サプライチェーン汚染、認可不足、監視不能」という構造的リスクが存在する。2025年の監査データでは、実装の43%で致命的な脆弱性が確認されており、大手企業においても情報漏洩事故が既に顕在化している。
このリスクにより、企業はサイバー攻撃の踏み台や情報漏洩によりガバナンス不全に陥る。結果として「AI不信」による全面禁止や過度な規制を招き、日本のAI導入とイノベーションが停滞する「エコシステムの崩壊」につながる甚大な影響が懸念される。

対策技術の評価

Tumiki MCP Managerは、AIとMCP間の通信を仲介する「プロキシ型ゲートウェイ」である。IDベースの階層的権限制御とリアルタイム不正検知により、既存防御をすり抜けるMCP通信を完全に統制する。

評価の結果、実用性と堅牢な防御が両立された。 追加遅延は平均163msでUXを維持し、権限昇格攻撃とサプライチェーン汚染攻撃は完全に阻止された。 全ての操作が証跡ログとして記録され、構造的なリスクに対し強固な防御層を提供する。

新規性・将来性

階層的認可で「許可機能のみ」の仮想MCPサーバーを構成。2段階防御（ホワイトリスト＋リアルタイム検知）で汚染リスクを排除し、実行者IDを紐付けた完全な証跡を保証。AI通信のゼロトラストを実現する。

公共性

本技術は特許・PCT出願済み。
ソースコードと評価データをGitHubで公開予定。およびクラウドサービスとしても提供。
これにより、安全なAI活用を推進し、イノベーション停滞リスクを防ぐ。

本事業の提案内容（株式会社レトリバ・株式会社万葉）

会社概要	株式会社レトリバ: 自然言語処理及び機械学習を用いたソフトウェアの研究・開発・販売・導入およびサポート 株式会社万葉: web系情報システムの構築・コンサルティング
提案名	LLMの脆弱性診断を行い、診断結果から改善用学習データを作成するWebアプリケーションの開発

特定したリスク

本提案では、業種ごとに想定される生成AIに関する懸念は異なるものと考える。その中で、我々は「本邦の生成AI導入者が想定する脆弱性に関して、生成AIが不適切な回答を生成するリスク」に着目した。通常の生成AI利用でも想定される「誹謗中傷」や「誤情報」といったものが一般的な脆弱性と考えられ中で、「社内システムと紐づいた不正行為」や「危険な作業手順の助長」などの分野に依存した脆弱性が想定される。本提案ではそのような、分野に応じた脆弱性に対して、生成AIが不適切な回答を行うことをリスクとみなした。

対策技術の評価

本提案では、特定したリスク緩和に向け、「LLMの脆弱性診断を行い、診断結果から改善用学習データを作成するWebアプリケーションの開発」を実施する。具体的には、生成AI導入者が想定した脆弱性に関する攻撃シナリオに基づいて、自動で攻撃プロンプトを作成し、対象LLMの脆弱性評価を行う。また、脆弱性評価結果と生成AI導入者が定義した防衛方針に基づき、対象LLMの改善用学習データを作成する。検証の結果、本技術を通して作成した学習データを用いたLLMは、脆弱性診断において改善することを確認した。

新規性・将来性

新規性として、生成AI導入者が定義した脆弱性診断と診断結果を用いた学習データ生成を組み合わせた点があげられる。課題として、現状では特殊な業種での評価が不足しているため、さらなるPoCなどでの検証が望まれる。

公共性

本技術はApache2.0で公開予定である。本技術の公開に伴い、生成AIが業種特有の脆弱性に対して不適切な応答を行うリスクを低減できる。これにより、様々な業種において生成AIを原因とした消費者被害や業務事故の抑制が期待できる。