BRiDGE

研究開発とSociety 5.0との橋渡しプログラム

programs for **B**ridging the gap between **R**&d and the **ID**eal society (society 5.0) and **G**enerating **E**conomic and social value

令和６年度補正予算 研究開発等計画

AI安全性強化に向けたLLM生成テキストの検証・分析・改善等支援技術の研究開発

令和７年１月

総務省

● 実施する重点課題（特に該当するものには◎、そのほかで該当するものがあれば○（複数可）を記載）

SIPや各省庁制度による研究開発成果の社会実装・市場開拓の加速化	他の戦略分野等との技術の融合による研究開発	スタートアップによるイノベーションの創出・促進	産学官を挙げた人材の育成・確保	グローバルな視点での連携強化
◎	◎			

● 関連するSIP課題（該当するものに○を記載）

持続可能なフードチェーン	ヘルスケア	包摂的コミュニティ	学び方・働き方	海洋安全保障	スマートエネルギー	サーキュラーエコノミー	防災ネットワーク	インフラマネジメント	モビリティプラットフォーム	人協調型ロボティクス	バーチャルエコノミー	先進的量子技術基盤	マテリアル事業化・育成エコ
			○										

1．施策の概要

- 今後、生成AI（大規模言語モデルLLM）が社会に浸透し、様々な業務（報告書、概要、要約、分析等の文書作成事務）においてLLMの利活用が進展することが想定される。一方、**国内外で開発されているLLMは内部の情報処理過程がブラックボックス化**しているため、生成するテキストについては、根拠のない情報、矛盾のある情報等を含む場合があり、LLMの出力が正しい内容であることを確認すること等により**LLM活用の安全性を確保するため、生成されたテキストの根拠・矛盾等の検証、改善等を支援する技術が必要**である。

- 本研究開発では、LLMが作成したテキストについて、政府・自治体が公開している情報など、内容の正確性が保証されている情報を活用し、**新たな AI技術（意味的・論理的関係を考慮した検索・質問応答技術の応用等）により、そのテキストの根拠・矛盾等を検証し、テキストに対する反論、改善等を支援する技術の研究開発を実施**する。

- 既存技術としてRAG（Retrieval Augmented Generation）と呼ばれる、生成前に検索を行ってヒントを得る技術があるが、生成の過程で誤った情報の混入が否定できない。また、統計データベースを用いて、生成されたテキストの根拠、矛盾等を確認する機能を開発するなどの取り組みがあるが、本研究開発は、Webテキスト等を活用して**より効果的かつ網羅的に対象とするテキストの根拠・矛盾の検証**をするほか、**テキストの確認だけでなく改善等を支援する技術であり、新たな取り組み**である。また、研究開発成果は、国内外を問わず様々な社が提供する生成AIが生成したテキストにも活用可能でることから**汎用性が高く、我が国における生成AIの安全性の向上に資する**。

- **SIP「ポストコロナ時代の学び方・働き方を実現するプラットフォームの構築」**においては、インターネット上の情報等を基に多様な意見をAIで生成し、様々な意見・価値に触れて学習できるオンライン・バーチャルツールを開発することで、**へき地等で十分な対話機会のない人々に十分な学習機会を提供する取組**を実施している。**生成AIの出力には、誤情報、根拠のない情報、矛盾した情報等が含まれ得るため、内容の正確性が保証された情報源を参照・活用し生成AIの安全性を高めることで、ポストコロナ時代の新たな学び方等の社会実装を一層加速**する。

2．施策の対象・成果イメージ

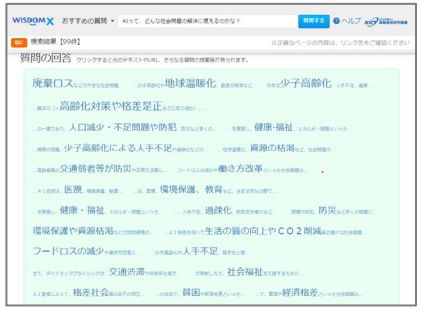
- 生成AIが生成したテキストの根拠・矛盾等の検証、改善等を支援する技術が確立。
- 成果は、民間等がライセンスにより活用可能とする予定であり、民間企業等が本成果を組み込んだ生成AIサービスを提供することが可能となる。これにより、民間企業等の生成AIの安全性を高め、社会全体のAI安全性を高めることに貢献。

3．資金の流れ

- WISDOM Xは、NICTが開発している、Web上のテキストを対象とする質問応答システムで、意味的・論理的関係を考慮した検索を得意とする。
- 既存の検索エンジンは多くの場合、質問の回答を網羅的に集めるためにはユーザが提示された文書を大量に読む必要があるが、WISDOM Xは、質問の端的な回答のリストを提示することができるため、関連する情報の全体像を迅速かつ容易に把握可能。

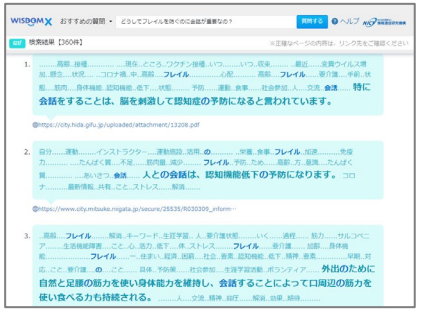
WISDOM Xの実行例

「なに？」に答える



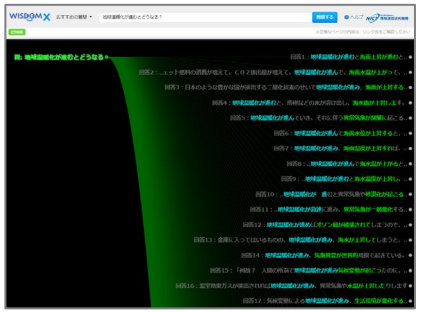
表現の同義性などを考慮

「なぜ？」に答える

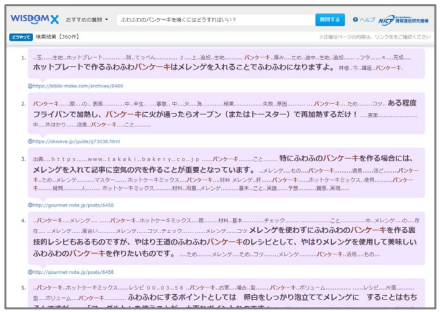


因果関係などを考慮

「どうなる？」に答える



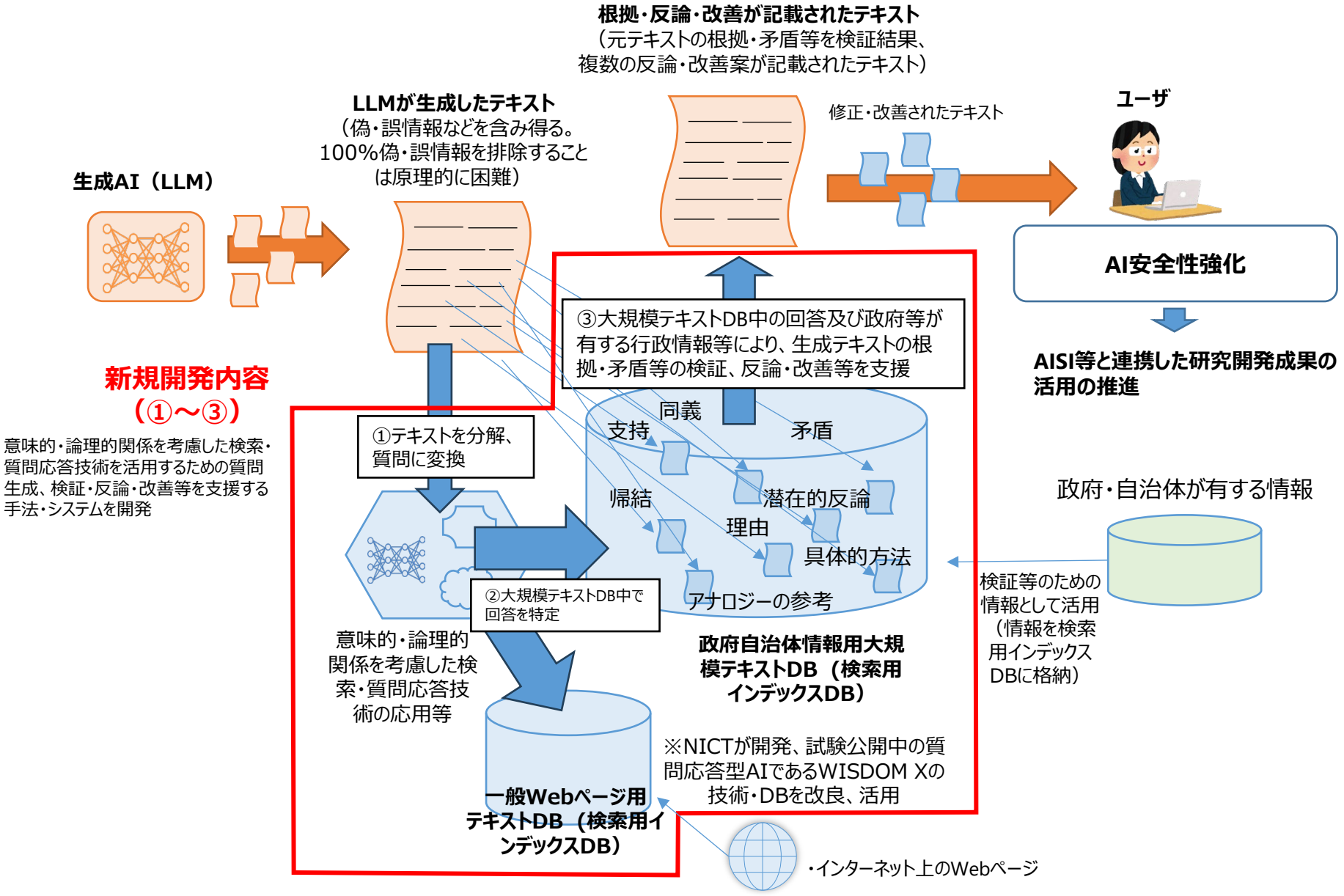
「どうやって？」に答える



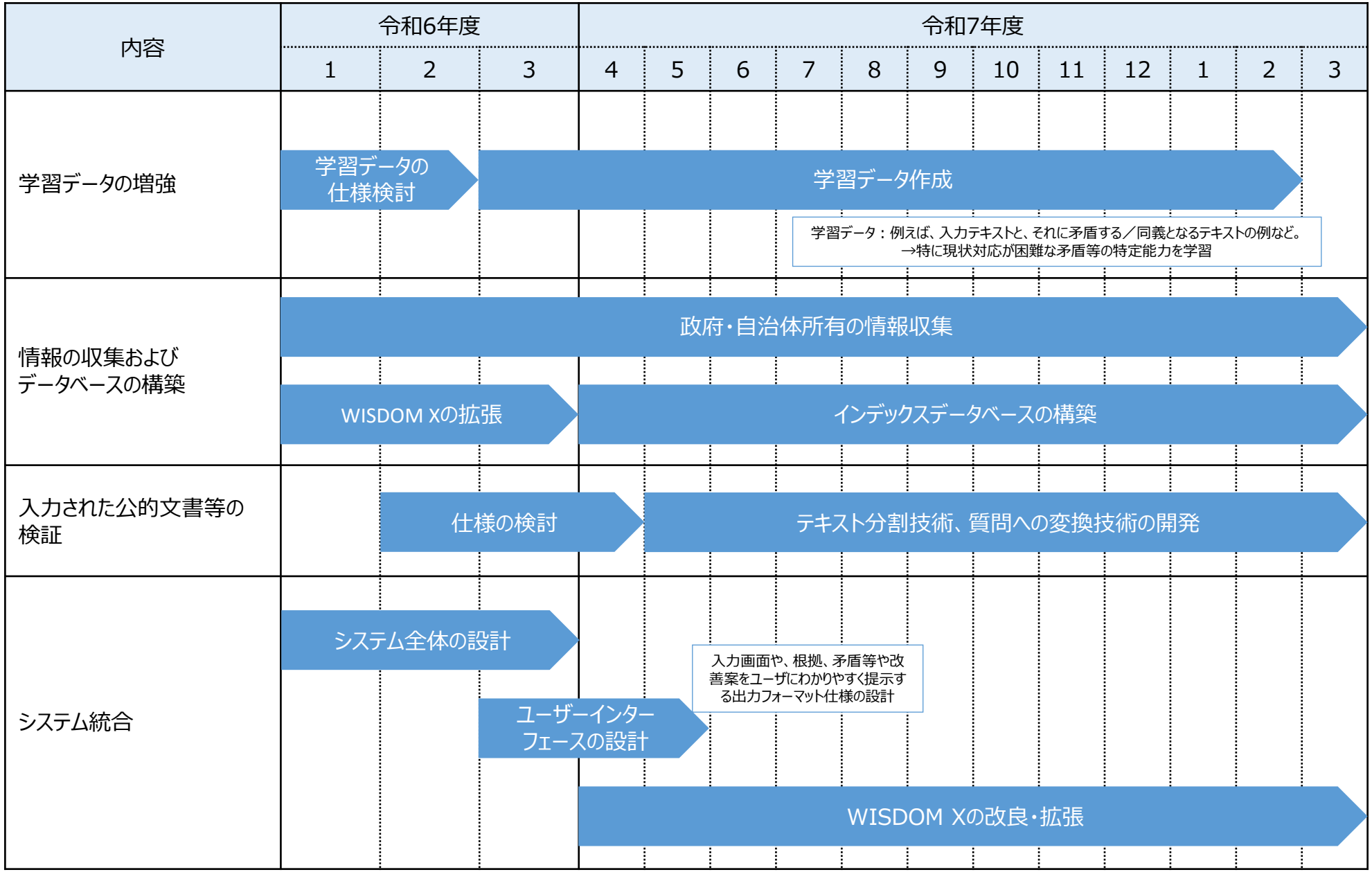
手段・方法の構造を考慮

（WISDOM X： https://www.wisdom-nict.jp/ ）

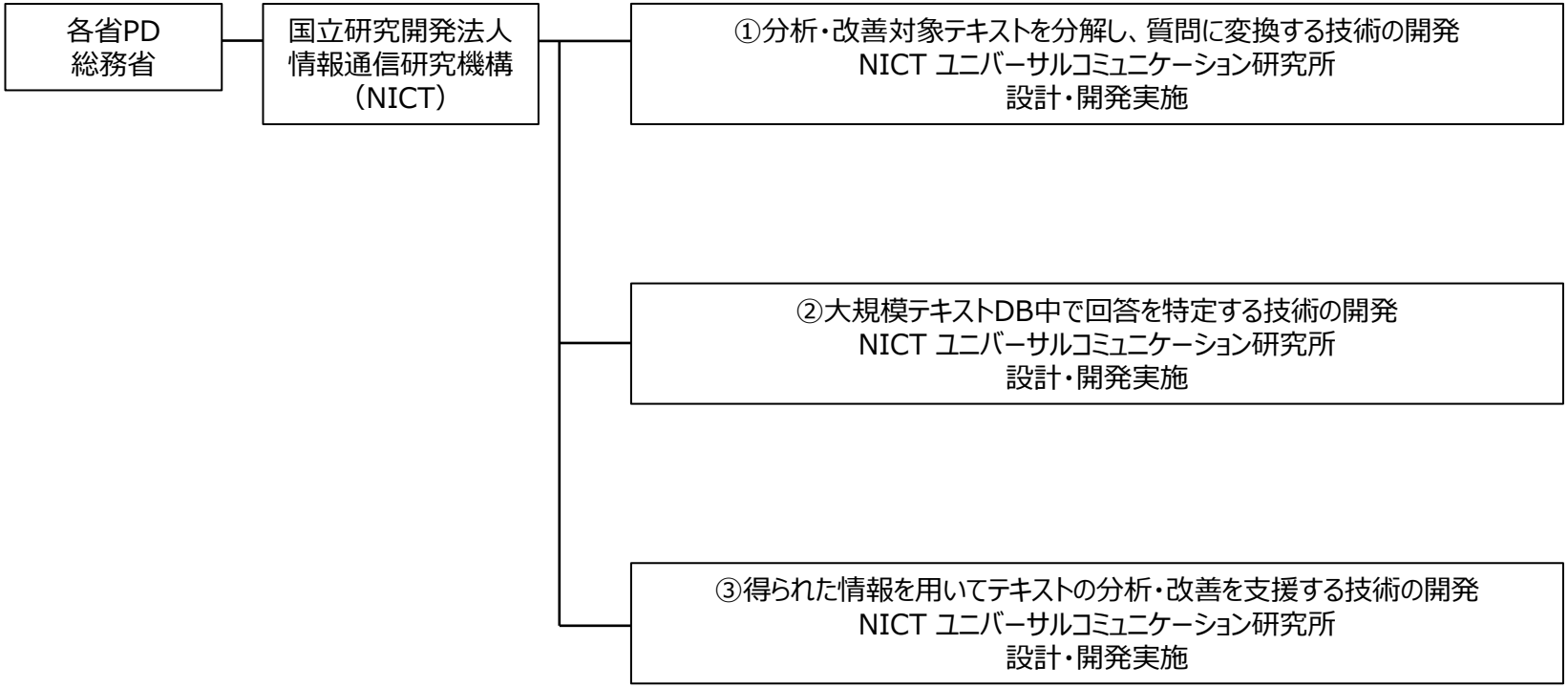
４．取組内容（システム概念図）



5．取組スケジュール


内容
令和6年度
令和7年度
1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
学習データの増強
学習データの仕様検討
学習データ作成
学習データ：例えば、入力テキストと、それに矛盾する／同義となるテキストの例など。
→特に現状対応が困難な矛盾等の特定能力を学習
情報の収集およびデータベースの構築
政府・自治体所有の情報収集
WISDOM Xの拡張
インデックスデータベースの構築
入力された公的文書等の検証
仕様の検討
テキスト分割技術、質問への変換技術の開発
システム統合
システム全体の設計
入力画面や、根拠、矛盾等や改善案をユーザにわかりやすく提示する出力フォーマット仕様の設計
ユーザーインターフェースの設計
WISDOM Xの改良・拡張

６．実施体制


各省PD
総務省
国立研究開発法人
情報通信研究機構
（NICT）
①分析・改善対象テキストを分解し、質問に変換する技術の開発
NICT ユニバーサルコミュニケーション研究所
設計・開発実施
②大規模テキストDB中で回答を特定する技術の開発
NICT ユニバーサルコミュニケーション研究所
設計・開発実施
③得られた情報を用いてテキストの分析・改善を支援する技術の開発
NICT ユニバーサルコミュニケーション研究所
設計・開発実施

7．実施内容・到達目標（KPI）

テーマ名	実施内容の概要 到達目標（KPI）
①分析・改善対象テキストを分解し、質問に変換する技術の開発	偽・誤情報などを含み得るLLM生成テキストに対する分析・改善を支援するために、対象テキストを分解し、分析・改善に必要な情報を得るための質問に変換する技術を開発。より具体的には生成テキストを分解した断片を、同義、支持、帰結、矛盾、理由、潜在的反論、具体的方法、アナロジーの参考等を特定するための質問に変換する技術を開発する。 到達目標：検証対象のテキストを分解して得られるテキスト断片(基本的には日本語の述語、その主語、目的語等からなる、いわゆる節に相当)のそれぞれに対して、開発するLLMによるテキスト生成やヒューリスティックな質問作成手法等を駆使し、各テキスト断片につき平均で16件以上の質問を作成する。
②大規模テキストDB中で回答を特定する技術の開発	LLM生成テキストに対して分析・改善を行うために必要な情報を得るため、質問応答型AIであるWISDOM Xの技術を改良、活用し、大規模テキストDB中から、①で開発する技術によって作成された各種質問への回答を特定する技術を開発。 到達目標：①で開発する技術によって対象テキストから作成した各質問に対して上位３件の回答の中に最低1件の適切な回答が含まれる精度が平均して９割以上となること。ただし、回答が得られない質問は、回答が得られないことをもってテキスト断片が低信頼であると判断可能なため、この精度の計算では考慮しない。
③得られた情報を用いてテキストの分析・改善を支援する技術の開発	②大規模テキストDB中で回答を特定する技術の出力を元に、テンプレートやLLM等を用いてテキストの分析結果、改善方法を指示するコメントを自動作成する技術を開発する。 到達目標：検証対象テキストの各テキスト断片の８割に対して妥当なコメントを自動作成する技術を開発する。