SIP第3期「統合型ヘルスケアシステムの構築における生成AIの活用」公開シンポジウム
2025年4月30日

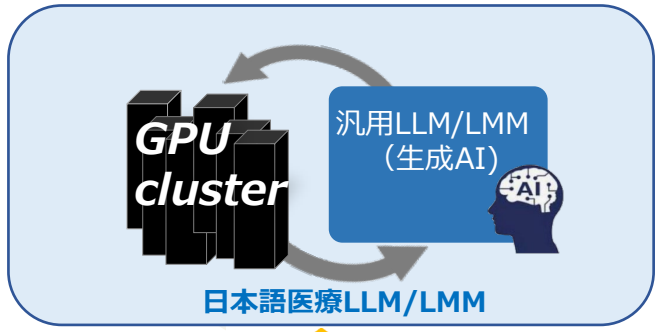
医療データ・医療 LLM/LMM の利活用を促進する医療データ基盤

合田　憲人
情報・システム研究機構

研究の目的

我が国の医療 LLM/LMM の研究開発を促進

- 医療データを収集・加工・蓄積し、研究者が利用可能とする**医療データ基盤**を構築
- 医療データ基盤上の**データを安全かつ適正に利活用するための仕組み**を考案



セキュリティ

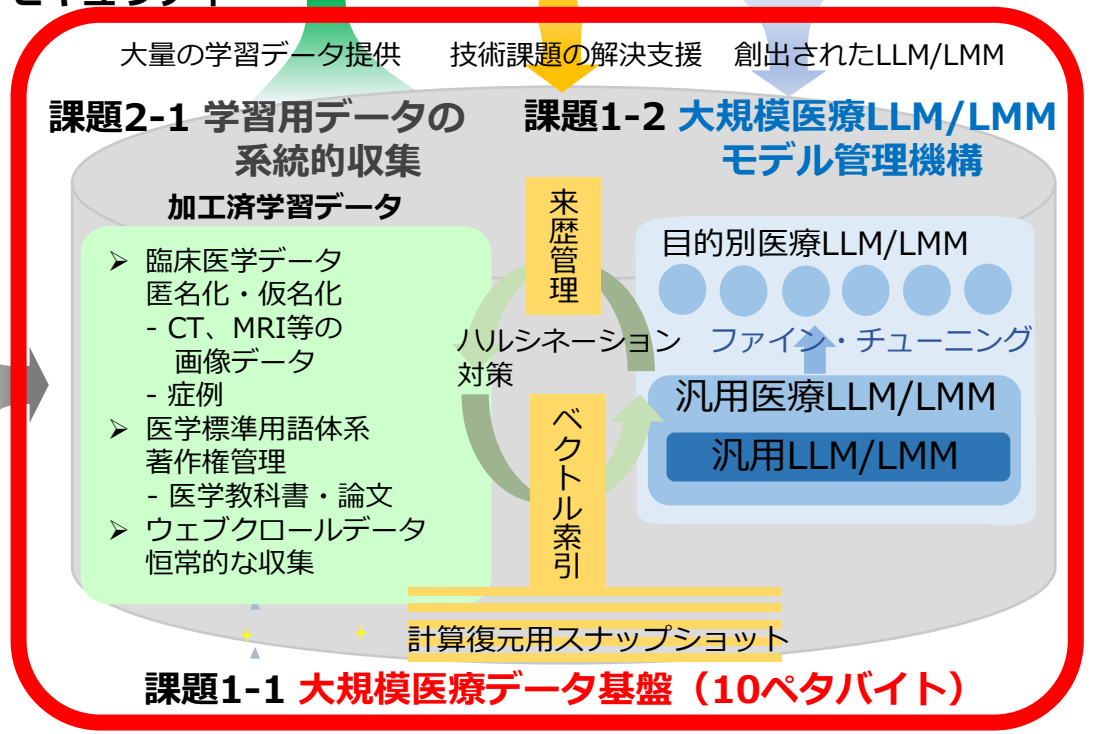
研究開発を促進するために解決すべき課題

1. 大規模医療データ基盤
   - 1-1 大規模医療**データの収集**・管理が困難
   - 1-2 多様な**LLM/LMMモデルの管理**が必須
     ハルシネーション等の**技術課題解決への支援**が必須
2. データ収集・持続的な利活用
   - 2-1 医療 LLM/LMM 開発のための系統的なデータ収集が困難
   - 2-2 医療データの**適正利用**のための仕組みがない
   - 2-3 **法・倫理的課題**の解決が必須

課題2-2, 2-3
医療データの適正利用管理

医療データの適正利用管理
- 医療データ・ガバナンス・データベース
- 適正利用ガバニングボード

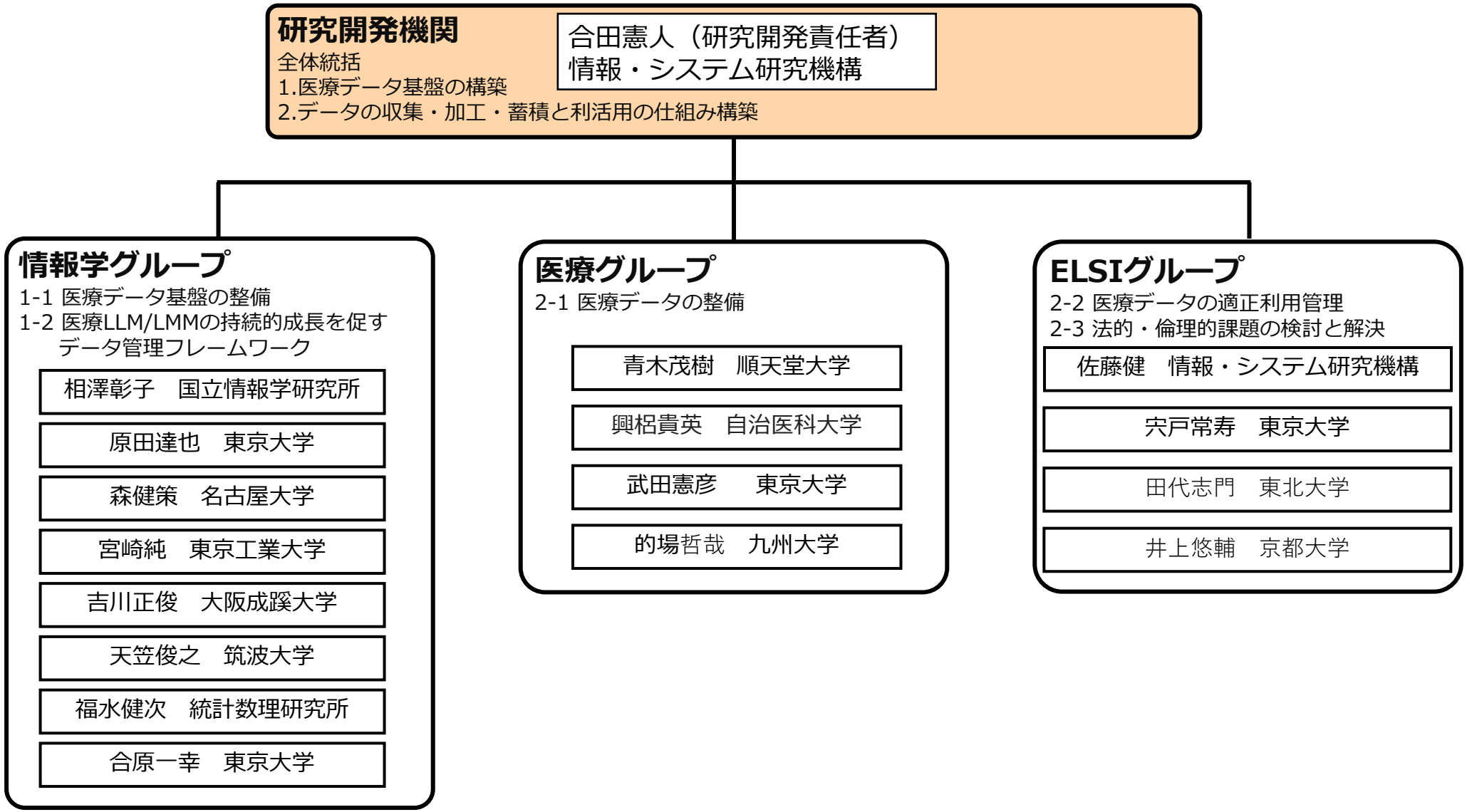
ELSIグループ
- 医学者、法・倫理学者、弁護士、IT研究者
- 著作権管理、個人情報保護、仮名化データ利用



医療データ・医療LLM/LMMの利活用を促進する大規模医療データ基盤

研究実施体制

情報学、医療、ELSIの専門家から構成される研究体制



医療データ基盤の利用イメージ

事例：テーマ2開発の汎用LMM（Asagi-2B, Asagi-4B, Asagi-14B）を格納し、プロジェクト内で先行リリース

- ✓ 学習データを迅速に取得、安心して利用できる
- ✓ 開発したモデルを効率よく共有・再利用できる

LLM/LMM 開発者

GPU GPU GPU GPU

③モデル開発（学習）
- **汎用モデル**の開発
- **医療モデル**の開発

②利用条件に従ったデータ取得

学習データ

モデル LLM/LMM

④モデル登録

ナイトセッション デモ

- ✓ モデルを検索、安心して利用できる
- ✓ モデルの来歴を調べられる

攻撃検知システム
- ✓ ストレージ不正操作防止
- ✓ 暗号通信にも対応

⑤利用条件に従ったモデル取得

医療データガバナンスDB

医療データDB　**モデル管理DB**

①データ・利用条件登録
SINET L2VPN経由の安全なデータ転送
またはHDD等直送

学習データ

モデル LLM/LMM

ベクトル索引

医療データ基盤

SINET L2VPN

データ提供者

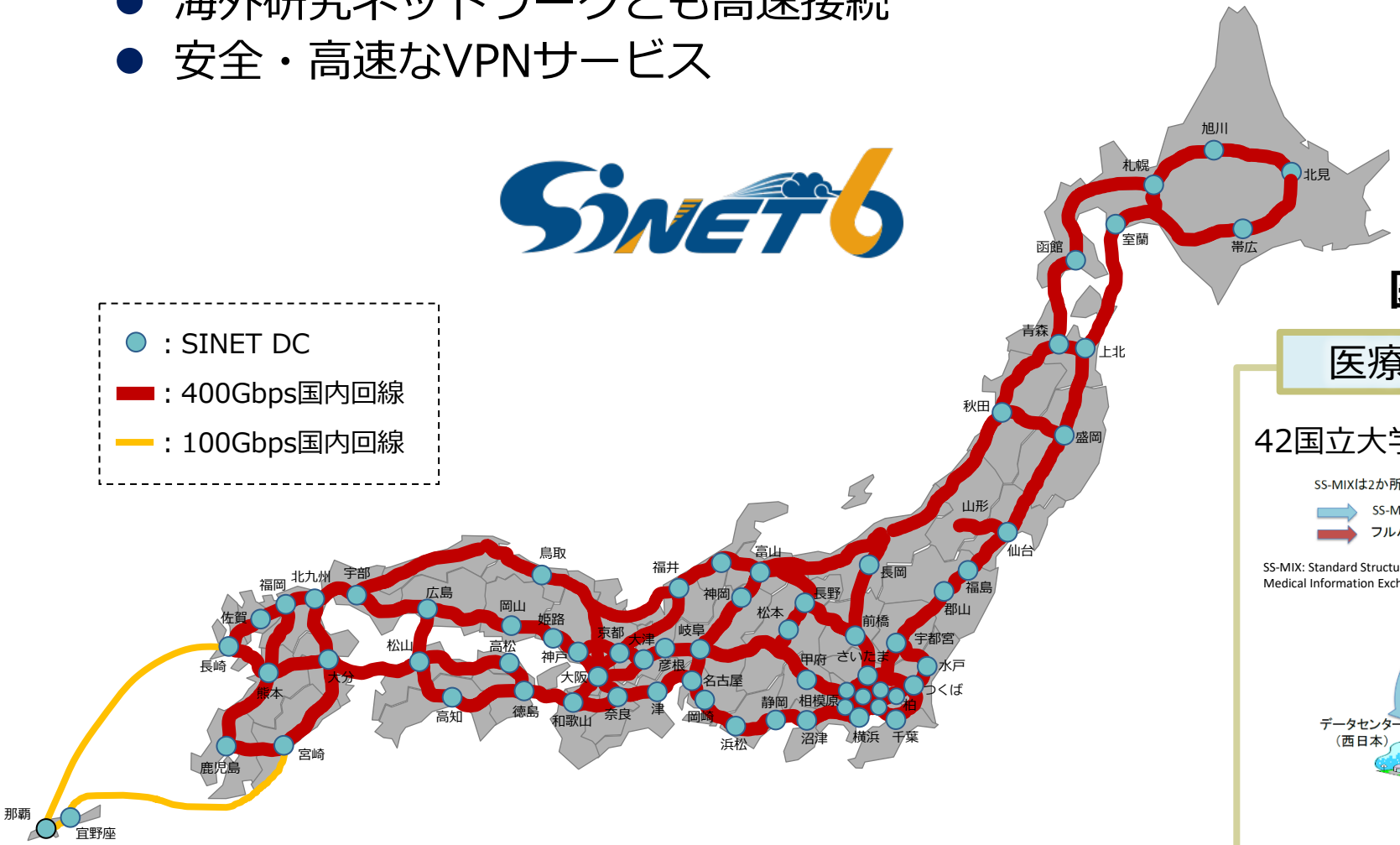
LLM/LMM 利用者

⑤モデルの利用
- **医療研究者**が取得したモデルをベースとして医療特化型モデルを開発
- **情報学研究者**がモデルを調査・改善（ハルシネーション対策等）

学術情報ネットワーク　SINET

国内の大学・研究機関、海外研究ネットワークを接続する学術情報ネットワーク

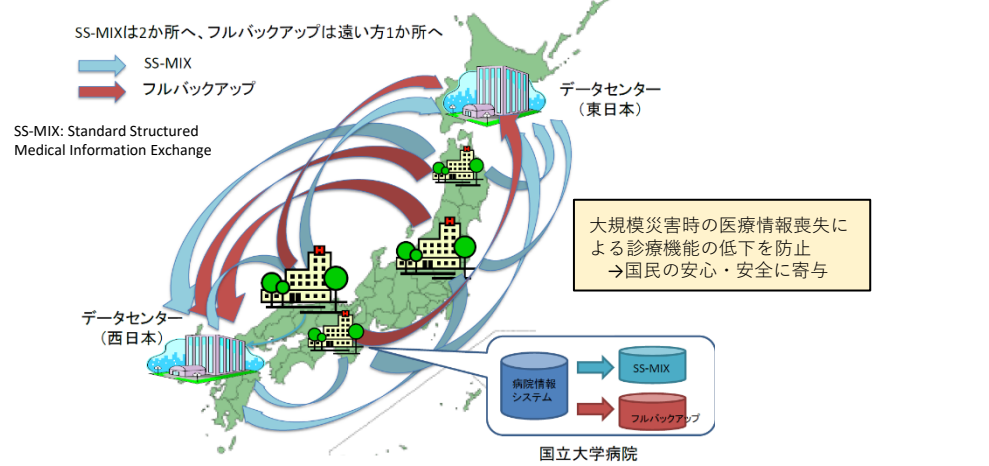
- 全国を400Gbpsのネットワークで接続（1,000以上の機関が加入 2025/3時点）
- 海外研究ネットワークとも高速接続
- 安全・高速なVPNサービス



医療分野でも多数の利用事例

医療情報バックアップ – 全国個別L2VPN

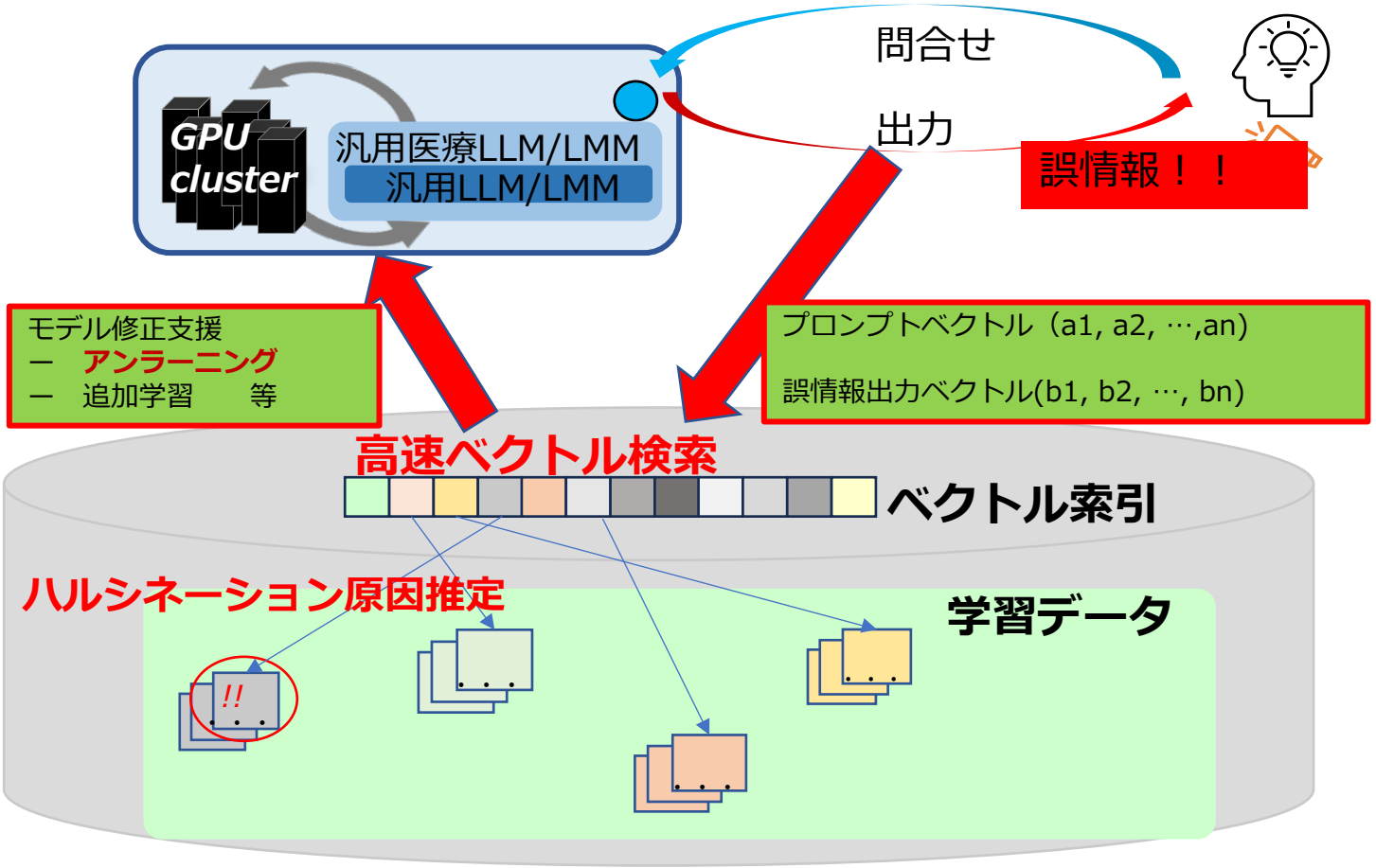
42国立大学46病院の医療情報を東西DCに定期的にバックアップ



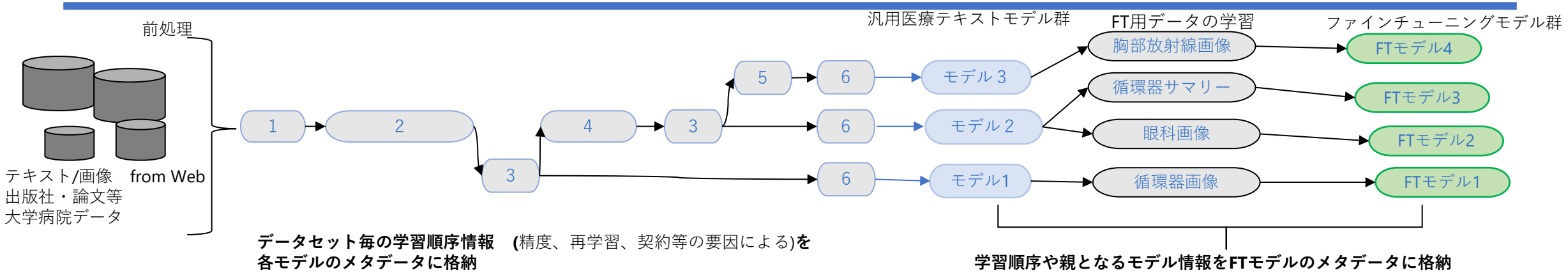
技術課題解決支援機構

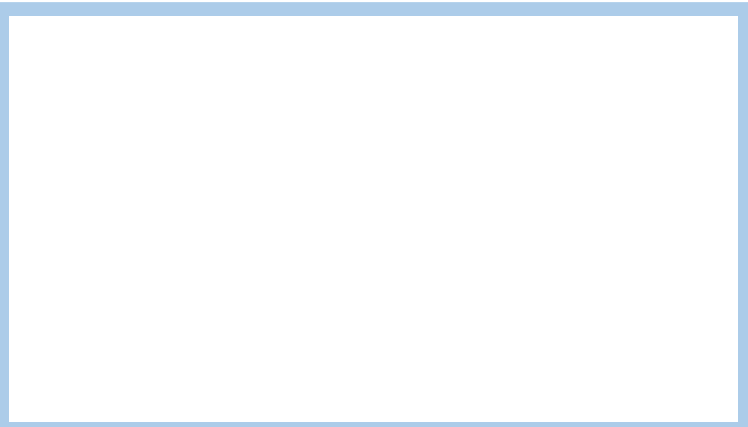
ハルシネーション対策、アンラーニングほかの**LLM特有の技術課題解決を支援するための機構**を提供

（例）ハルシネーション発生時の原因究明を支援する機構

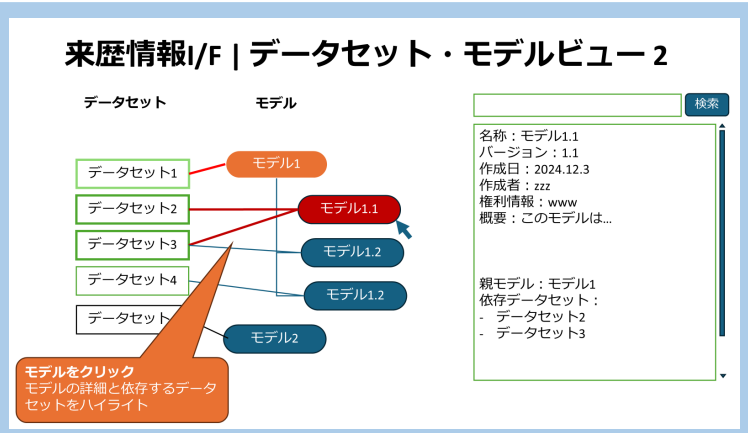


技術課題解決支援のための研究開発：来歴管理

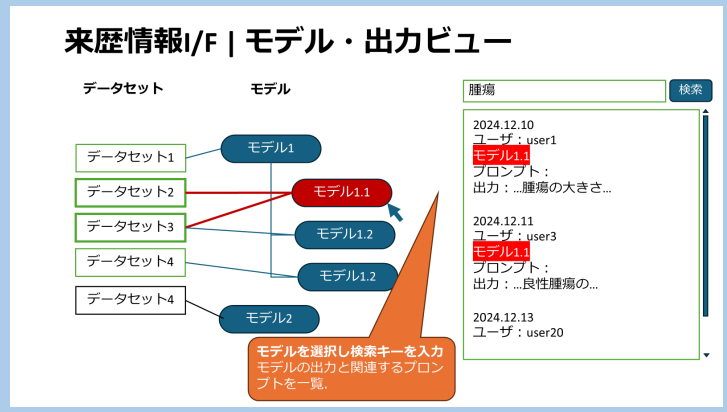




データセットを起点に利用モデルを提示



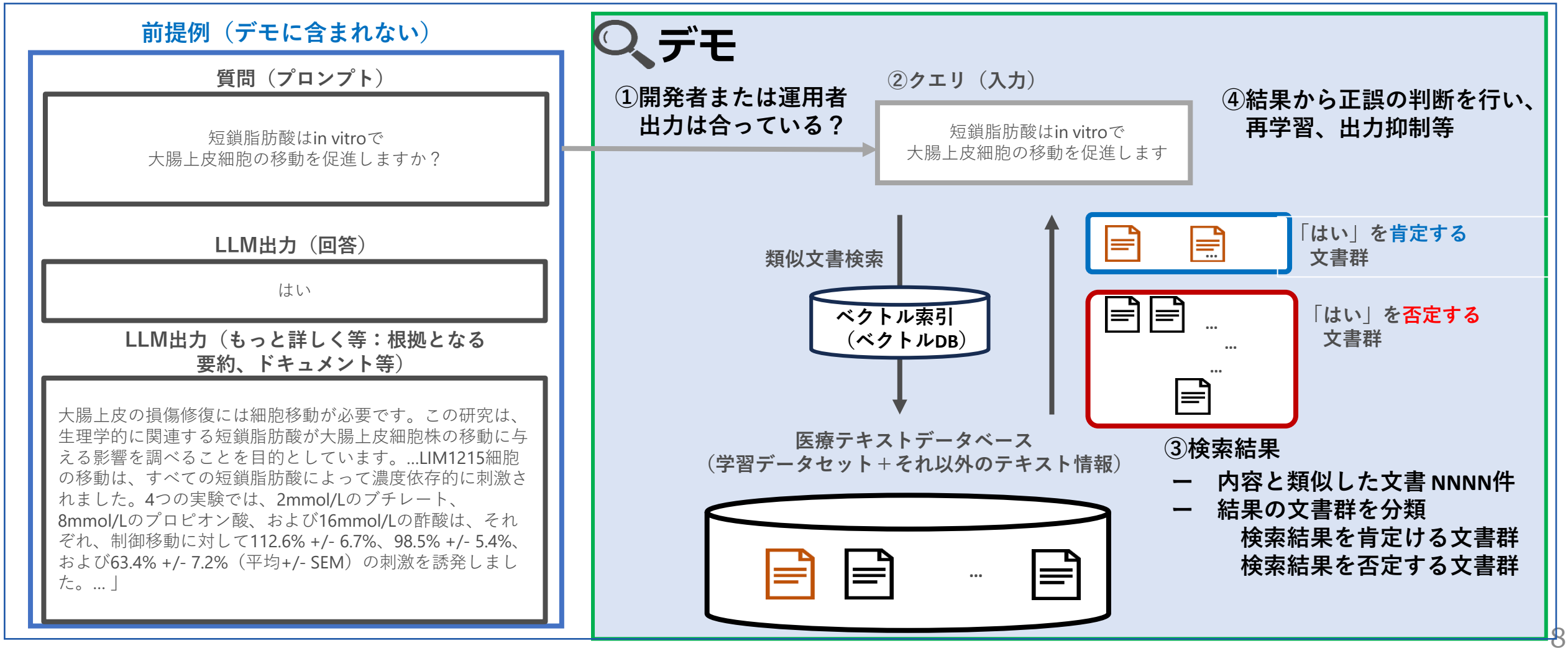
モデルの来歴を提示



モデルの動作例などを提示

技術課題解決支援機構：デモンストレーション例

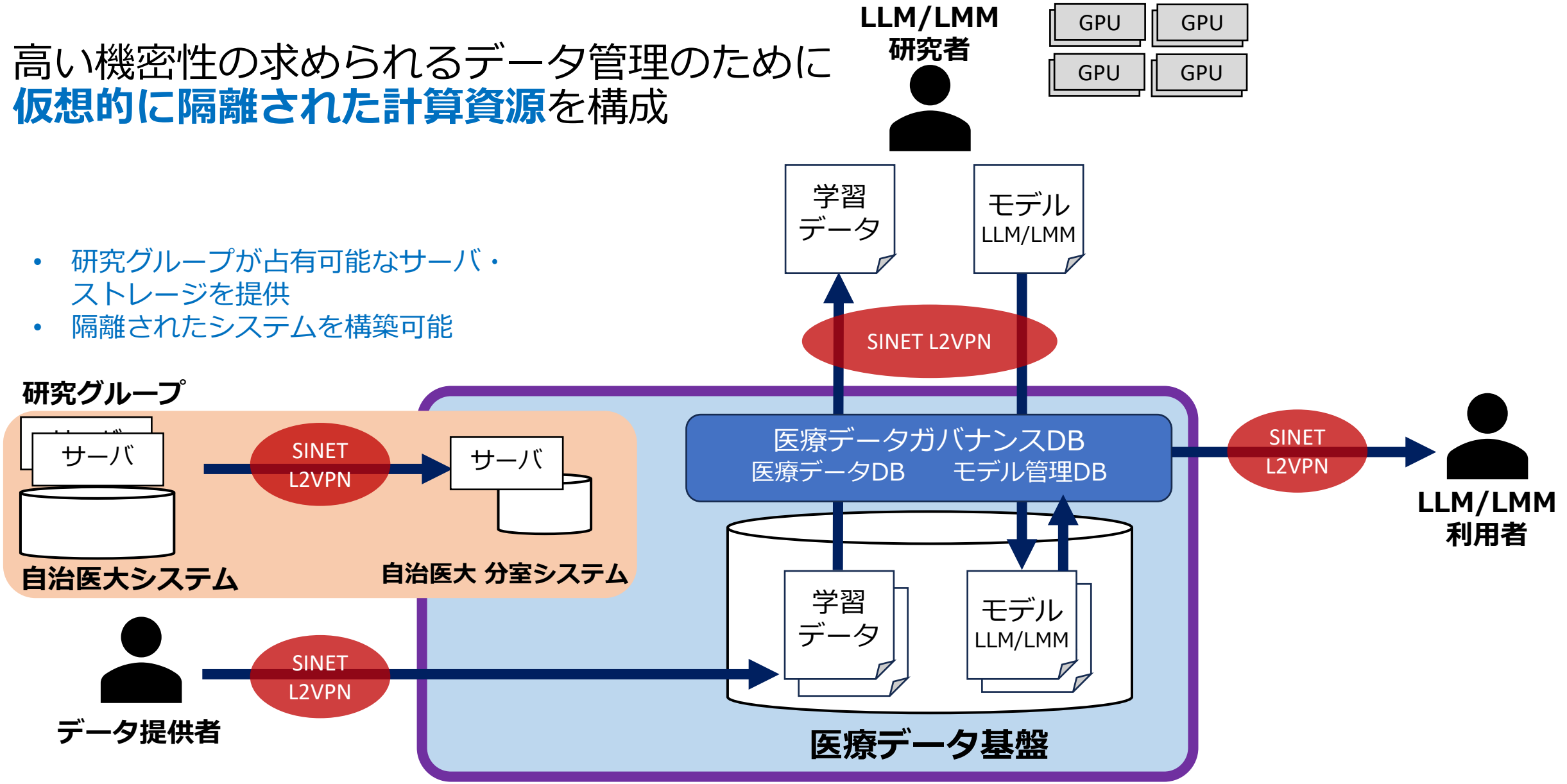
ハルシネーション対策支援機能　誤りに関連する文書抽出・文書の支持/不支持表示



技術課題解決支援のための研究開発：医療用仮想プライベートクラウド

高い機密性の求められるデータ管理のために
仮想的に隔離された計算資源を構成

- 研究グループが占有可能なサーバ・ストレージを提供
- 隔離されたシステムを構築可能



医療データの整備

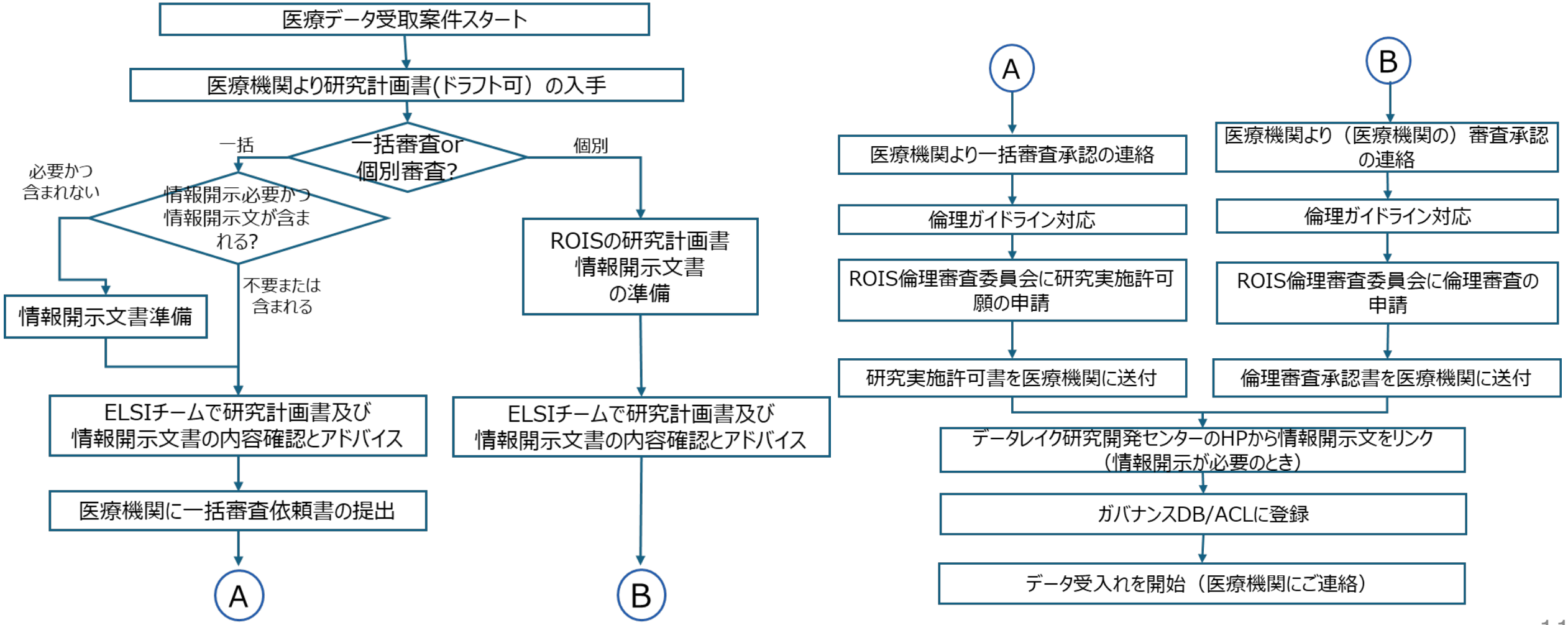
医療データ基盤に格納予定のデータについて、各データの概要、倫理・契約条件、優先度を整理し、策定した**データ受入れプロセスに従って安全で円滑なデータ受入を実現**

医療データ基盤で管理するデータ一覧（テーマ１，２，４全体）

データカテゴリ（分類）		件数・量
クロールデータ		
	テキスト	2.12兆トークン
	画像・テキストペア	18億ペア
高品質医療テキスト		
	医学書院教科書・医中誌Web	17.6億文字
医療データ		
	SIPテーマA-1参加13施設(CLIDAS)からのデータ等（DPCレセプト、心電図・心エコー数値、胸部X線画像）	3.8万件
	日本医用画像データベースのCT・MR画像	5.6億枚
	特化型LLM/LMM学習用データ	119万件

医療データ受入プロセス

倫理ガイドラインに沿った**データ受入のためのプロセスをELSIの支援を得て策定**し実践

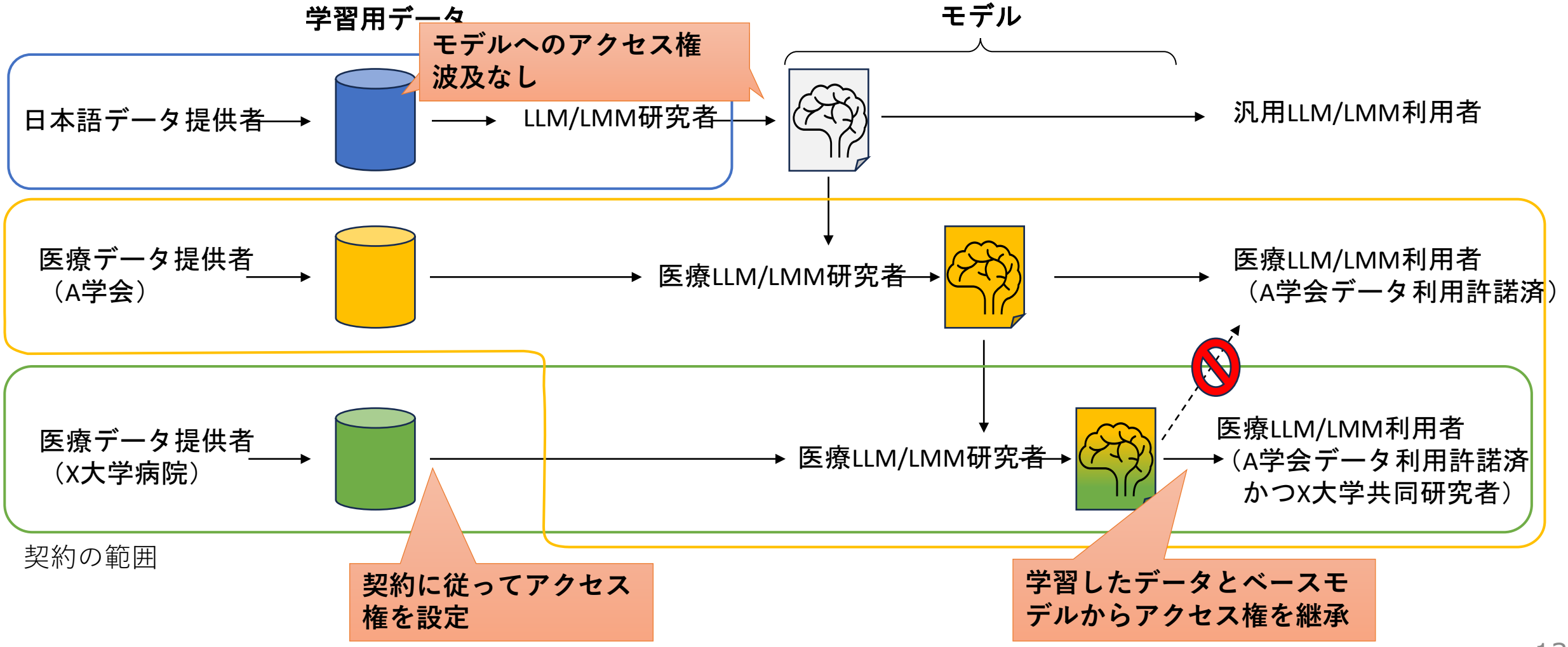


医療データの例

データ概要					データ仕様				工程等			備考
ID	データ名	内容	SIP テーマ	研究機関（担当）	ファイル数	データフォーマット	総データサイズ（GB）	加工状態	提供機関倫理委員会	ROIS内倫理審査	データ移行開始(予定)	研究計画書名
51	NII医療ビッグデータ研究センター放射線画像データベース	日本医学放射線学会の主催するJ-MIDからNII医療ビッグデータ研究センターに提供された画像CT及びMRのDICOM画像及び所見文	テーマ２	順天堂大学 NII	5.6億枚	DICOM / NIfT / テキスト	300000	匿名化＋付加情報	2024/12	2024/12	2025/4	「日本医用画像データベースのナショナルデータベース化と画像診断支援技術開発・臨床応用に関する研究」
56	眼底写真50万枚	自治医科大学健診センター受診者の20年分の眼底写真	テーマ２	自治医科大	50万枚	JPEG	250	仮名化相当	2025/2	2025/2	2025/4	「医用画像の機械学習用データベース構築」
58	心電図データ	東大病院の心電図データ	テーマ２	東京大学	10万	CSV	30	仮名化相当	2024/11	2024/12	2025/2	統合型ヘルスケアシステムの構築における生成AIの活用
62	CT画像＋レポートペア-CT画像	がん疾患の体幹部CT画像（DICOM）	テーマ２	自治医科大	3000件	DICOM /NIFTI	2000	仮名化相当	2024/12	2025/1	2025/4	「自治医科大版DataLake]
65	心電図データ(float形式数値csvデータ 12*5000)	自治医科大学健診センターでの過去の検査歴	テーマ２	自治医科大	17万	CSV	20	仮名化相当	2025/3	2025/4	2025/5	機械学習手法を利用した心電図検査による心血管イベント予測能の検討
71	CLIDAS-DPC	DPC入院レセプト 10万症例、検体検査、処方データ	テーマ４	九州大学	10万症例	CSV	1	仮名化相当	2025/3	2025/3	2025/6	循環器疾患レジストリ研究(臨床効果データベース整備事業　CLIDAS研究)
72	CLIDAS-ECG	心電図数値＋MFER	テーマ４	東京大学	50万件	CSV/MFER	1	仮名化相当	2025/3	2025/3	2025/6	循環器疾患レジストリ研究(臨床効果データベース整備事業　CLIDAS研究)
74	CLIDAS-XP	胸部X線画像	テーマ４	東京大学	30万件	DICOM	15000	仮名化相当	2025/3	2025/3	2025/6	循環器疾患レジストリ研究(臨床効果データベース整備事業　CLIDAS研究)

医療データ適正利用管理

医療データを契約条件に基づいて管理する**ガバナンスDBとアクセス制御機能**により、医療データ及び学習モデルの適正な利活用を実現



ELSI グループ設置による法的・倫理的課題の検討

検討結果や提言を報告書（７００ページ）として**公開予定**

https://ds.rois.ac.jp/center8/

検討事項	成果
生成AI開発フェーズでの法的・倫理的課題の検討	● 法・倫理指針に則ったデータ収集・提供の枠組みを整理・実運用 • 著作権法：非享受目的での利用・著作権者との個別契約 • 個人情報保護法：学術研究例外の適用 • 人を対象とする生命科学・医学系研究に関する倫理指針
研究途上において生起する課題の検討とフィードバック	● テーマ１・２の研究者への法・倫理に関する支援：チュートリアル、アドバイス ● 医療テキスト出版社・団体との契約支援 ● データ受け入れ手順の作成・運用支援
生成AI利活用フェーズでの法的・倫理的課題の検討	● 著作権法・個人情報保護法：法的視点からの侵害を防ぐ技術の必要性を開発へフィードバック ● 仮名加工情報、社会実装を前提とした法的枠組み整理（次世代医療基盤法、共同利用）
諸外国の制度の整備状況についての情報収集と予想される国内法規制への対応検討	● 海外調査：各国の医療生成AIに関する制度・動向調査（EU、米、デンマーク等） ● 政策提言の作成：医療データ活用、安全性確保、AIガバナンスの視点等から提言を整理し、報告書と共に公開

まとめ

LLM／LMM研究者が医療データ基盤を利用することにより：

貢献	貢献につながるサブテーマ
医療データ基盤上の大量のデータを学習データとして利用可能	1-1 医療データ基盤の構築 2-1 医療データ整備
作成したモデルの保存・管理、先進技術支援機能（ベクトル索引等）を利用して、モデルの利活用、LLM/LMM特有の技術課題（ハルシネーション対策等）を解決するための研究を推進可能	1-1 医療データ基盤の構築 1-2 データ管理フレームワーク
医療データの利用許諾条件や関連法等に基づいた医療データの適正な利用が可能	2-2 医療データの適正利用 2-3 ELSIグループ設置による法的・倫理的課題の検討