

# AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第4回） 議事録

[https://www.mext.go.jp/b\\_menu/shingi/gijyutu/gijyutu29/006/gijiroku/mext\\_00004.html](https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu29/006/gijiroku/mext_00004.html)

記事ページ本文

現在位置

トップ

>

政策・審議会

>

審議会情報

>

科学技術・学術審議会

>

情報委員会

>

AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ

> AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第4回）議事録

AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ（第4回）議事録

1. 日時

令和8年4月24日（金曜日）16時00分～18時00分

2. 場所

文部科学省東館17階局4会議室 及び オンラインのハイブリッド形式

3. 議題

AI for Scienceを支える研究データの管理・利活用及び流通の在り方について

その他

4. 出席者

委員

尾上主査、石田委員、江村委員、工藤委員、千葉委員、林委員、宮田委員、矢守委員、吉田委員、若目田委員

文部科学省

麻沼参事官補佐、池田参事官補佐、鈴木潤科学官、込山学術調査官、國本学術調査官

オブザーバー

国立情報学研究所

所長 黒橋 禎夫

副所長/アーキテクチャ科学研究系 教授 合田 憲人

副所長/情報社会相関研究系 教授 佐野 多紀子

情報学プリンシプル研究系 教授 武田 英明

情報社会相関研究系 教授 片岡 洋

オープンサイエンス基盤研究センター長 谷藤 幹子

アーキテクチャ科学研究系 教授 竹房 あつ子

知識コンテンツ科学研究センター 上級学術基盤研究員 中島 律子

5. 議事録

【尾上主査】 それでは、定刻になりましたので、科学技術・学術審議会情報委員会AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループの第4回会合を開催いたします。

委員の皆様におかれましては、お忙しいところお集まりいただきまして、ありがとうございます。本日は、現地出席とオンライン出席のハイブリッドでの開催としております。また、通信状態等に不具合が生じるなど続行できなかった場合、委員会を中断する可能性がありますので、あらかじめ御了承ください。

まず、事務局より本日の出欠状況などについて御案内願います。あわせて、事務局に異動があったとのことですので、御紹介願います。

【麻沼参事官補佐】事務局でございます。本日の出席者につきましては、10名全ての委員に御出席いただいております。

また、本ワーキンググループには、国立情報学研究所からもオブザーバーとして御参画いただいておりますが、本日は黒橋所長、合田副所長、佐野副所長、武田先生が現地から御参加いただいております。谷藤先生、竹房先生がオンラインから御出席いただいております。よろしく願いいたします。

続きまして、事務局及び陪席いたします科学官、学術調査官に関しまして、4月1日付けで異動がございました。

学術基盤整備室長に着任いたしました山本でございます。

【山本学術基盤整備室長】よろしく願いいたします。

【麻沼参事官補佐】総括の参事官補佐に着任しております池田でございますが、オンラインからの参加でございます。

新たに科学官として着任していただいております鈴木潤科学官でございます。

【鈴木科学官】よろしく願いいたします。

すみません。本日は最初の20分ぐらいで、その後、講義に行きますので、最初の20分ぐらい参加させていただきます。

【麻沼参事官補佐】よろしく願いいたします。

【鈴木科学官】よろしく願いします。

【麻沼参事官補佐】新たに学術調査官として着任しております國本千裕学術調査官でございます。

【國本学術調査官】國本です。よろしく願いいたします。

【麻沼参事官補佐】込山学術調査官につきましては、前期に引き続き着任していただいておりますので、よろしく願いいたします。

【込山学術調査官】よろしく願いします。

【麻沼参事官補佐】以上でございます。

【尾上主査】ありがとうございました。

次に、配付資料の確認とハイブリッド開催に当たっての注意事項について、事務局より御説明をお願いいたします。

【麻沼参事官補佐】事務局でございます。

それでは、議事次第に基づきまして、配付資料を確認させていただきます。現地出席の先生方は手元の配付資料を、オンライン出席の先生方はダウンロードいただいている資料を御確認をお願いいたします。

議事次第に配付資料一覧がございまして、資料1、2、3、参考資料1、2、3でございます。もし抜け漏れ等ございましたら、議事の途中でも結構ですのでお知らせいただければと思います。現地出席の先生方は手を挙げてお知らせいただくのと、オンライン出席の先生方は事務局までお電話で御連絡いただければと思います。

続きまして、ハイブリッド開催に当たって注意事項を申し上げます。

御発言時を除きまして、マイクは常にミュートとしていただくようお願いいたします。

ビデオは常にオンにいただき、通信状況が悪化した場合にはビデオを停止していただくようお願いいたします。

運営の都合上、現地出席の方も含めまして、御発言いただく際は挙手ボタンを押して御連絡をお願いいたします。

尾上主査におかれましては、参加者一覧を常に開いていただきまして、手のアイコンが表示されている委員の御指名をお願いいたします。

議事録作成のため、速記の方に入っておりますので、御発言される際は、お名前をおっしゃってから御発言をお願いいたします。

恐れ入りますが、マイクの数に限られておりますので、現地出席の先生方が御発言される場合には、大きめの声で御発言をお願いいたします。

また、本日もですが、傍聴希望をいただいた方は、YouTube配信により御参加いただいております。最後になりますが、トラブルが発生した際は、現地出席の先生方は手を挙げていただき、オンライン出席の先生方は電話にて事務局まで御連絡をお願いいたします。

事務局からの御連絡は以上でございます。

【尾上主査】ありがとうございました。

本日は、前回に引き続きまして、AI for Scienceを支える研究データの管理・利活用及び流通の在り方について、その他の2件の議題を予定しております。

前回は、AI for Science推進に当たっての基本的な戦略方針の検討状況を御紹介いただき、研究データの利活用を進める上での取扱いや考え方など、皆様からも御意見を頂戴いたしました。NII様からは、研究データ基盤をどのように構築してきたのか、事業推進や外的要因などから浮き彫りになってきている課題を整理しつつ、今後の取組方針などを伺いました。より高い視点での研究データ基盤、NIIの役割や位置づけを整理する必要があるとの御指摘も委員からいただいたところでございます。今回は、まず文部科学省より、第7期科学技術・イノベーション基本計画や3月31日に取りまとめられた「AI for Scienceの推進に向けた基本的な戦略方針」に関して、情報基盤関連の記載を中心に改めて御紹介いただきつつ、これまでのワーキンググループでの議論を少し振り返りながら、取りまとめに向けての整理を始めていきたいと思っております。

さらに、NIIからは、「学術研究プラットフォームの将来像」として、これまでに流通基盤SINET、研究データ基盤NII RDCの高度化について御説明いただき、議論を深めているところでございますが、今回は、研究者にとって、AI for Scienceの推進、研究データ基盤の利活用の肝となるであろう「知識基盤の構築」に関して、黒橋所長に御説明いただきます。

まずは、文部科学省から資料1に基づき御説明いただき、10分程度意見交換した後、黒橋所長より御説明いただきます。

では、麻沼補佐、どうぞよろしくお願いいたします。

【麻沼参事官補佐】ありがとうございます。

それでは、資料1を御覧ください。

1ページおめくりいただきまして、2ページ目を御覧ください。こちらがAI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループの取りまとめに向けてという資料でございますが、1ページ目、本ワーキンググループの取りまとめに向けたスケジュールを再度お示ししているものでございます。昨年度3回開催しておりまして、本日は、真ん中の4回目でございます。知識基盤に関して本日は御議論いただく予定でございます。その後、6月をめどに報告書を取りまとめさせていただきますと考えております。

そして、青いところですが、次期SINETにつきましては、2028年度から運用開始を目指しております。研究データ基盤の高度化を進めるべく、こちらは2027年度より新規事業の開始を目指しているという流れで動いております。

次のページ、3ページ目を御覧ください。こちらは第1回目にもお示ししております本ワーキンググループでの取りまとめの骨子案でございます。足りない観点等がございましたら随時追加していただければと思いますので、御意見等いただければと考えております。

続きまして、4ページ目以降ですが、これまでのワーキンググループでの主なポイントをまとめております。

まず5ページ目ですが、こちら第1回目のワーキンググループでの主なポイントとしてコメントをまとめさせていただきます。第1回目では、NIIがこれまで構築を進めてこられました「学術研究プラットフォーム」の現在地と今後の展開について御説明をいただきまして、委員の先生方からは、例えば、SINETにつきましては、AIによって研究プロセス自体が破壊的に変わる可能性があり、データ流通基盤も「従来型の増強」では不十分となる可能性を視野に入れるべきではないかといった御意見や、研究データ基盤NII RDCにおいては、研究データ基盤は単なるITではなく、研究者の作業負担軽減、時間創出イコール研究の質向上に直結するもの、また、個別大学・研究者がそれぞれ調達・交渉する構造は非効率的であり、調達・管理・交渉の集約化余地が大きいのではというご意見をいただきました。

また、セキュリティに関しても、個人情報・医療データ・社会データ等の機微データを扱うには「安心して置ける／流せる」基盤が不可欠であることや、AI for Scienceの対応についても、分野固有のデータ基盤を持つ外部のプラットフォームとどうつながるのかなど、他分野・他機関の基盤とどう連携するのかを明確化すべきですとか、最後、研究データの扱い等については、個人情報保護法の取扱いを例に挙げまして、学術研究用途は、一般的な民間利用よりも社会的受容性が高いと考えられるの

で、「個人データは使わない」という前提ではなく、どうすれば適切に使えるかといった方向で議論すべきといった御意見をいただいております。

続きまして、6ページ目が、第2回目の主なポイントになります。第2回目では、SINETの高度化、セキュリティ強化に向けた構想についてNIIより御説明をいただきまして御議論いただいたところですけれども、6ページ目がNIIの構想のポイントをまとめたものになりますので、本日は説明をは割愛させていただきますまして、7ページ目を御覧ください。こちらは委員のコメントをまとめたものです。

SINETの高度化につきましては、ネットワーク構成検討に当たり、必要な容量という観点だけではなく、世界最先端の研究実現を念頭に置いて検討してほしいや、データ容量がどこまで伸びるのかは予想するのは難しいだろうといった御意見もいただいております。

また、ネットワーク構築においては、ネットワークの整備計画は、大型実験施設の整備計画やコンピューティング基盤の整備と連携して考えたほうがよいですとか、ネットワークの運用・開発を行う人材不足についても言及いただいております。

セキュリティの観点からは、AIの利活用によって生じる、サイバーセキュリティや情報インシデント等への対策についても検討すべきであるとか、認知度の向上についても努めたほうがよいといったコメントもいただいております。

続きまして、8ページ目を御覧ください。こちらが第3回目での主なポイントになります。第3回目は、研究データ基盤の高度化、認証強化に向けた構想についてNIIより御説明をいただいております。8ページ目はそのポイントをまとめたものですので、割愛させていただきますまして、9ページ目を御覧ください。

9ページ目は委員のコメントをまとめたものになりまして、NII RDCのAI for Science推進と研究データ基盤につきましては、研究データ基盤も、単にAIのためのデータ置場になるのではなく、実験・現場・研究者にどう寄与するのが重要であることや、いろいろな分野からデータを吸い上げること自体が容易ではないといったことから、そもそもデータをどう創出・収集するのかという点にももっと重点を置く必要があるのではないかといただいた御意見をいただいております。

また、認証とアクセスの確保については、人文系の資料が大量に眠っていますけれども、それらの多くがまだデジタル化されていない、紙であるがゆえに朽ちてしまう可能性もあるといったことや、今後は、学術研究が持っている一次データが非常に重要になってくるのではないかといただいたコメントをいただいております。あわせまして、医学系や病院のデータについては、外に出せないデータが多いけれども、非常に価値が高いといったコメントもいただきました。

その後、総論のところ、一番最後ですが、単にツールやインフラを提供するだけでは不十分で、新しい人が入ってきてシステムを使おうとしたとき、問題が起きたときに分からないことが出てきたときなど、相談できる人的サポートがあると活用が進むのではないかといただいたコメントもいただいたところでございます。

続きまして、政策文書における関連記載を紹介させていただきます。

11ページ目になりますけれども、こちらが第7期科学技術・イノベーション基本計画でございます。3月27日に閣議決定がなされております。11ページ目は基本的考え方になりますので、本日は割愛させていただきますまして、12ページ目を御覧ください。

第2章の知の基盤としての「科学の再興」という項目になり、こちらの4ポツ目として、AI for Scienceによる科学研究の革新という項目が入れられております。

こちらが一番本ワーキンググループに関連する内容が書き込まれたところになっておりまして、例えば、(2) AI駆動型研究を支えるデータ創出・活用基盤の整備のところでしたら、太字になっているところになります。電子化されていないデータやレガシーデータの利活用などについても検討するとか、AI for Scienceの推進においては、オープン・アンド・クローズ戦略の下で、研究データの管理・利活用を推進する。その上で、日本の持つ研究データがAIの活用等により意図せず流出し、我が国の国際優位性などが損なわれることを防ぐため、研究データの扱いについては、AI技術の進展や研究分野・データの特性等に留意しつつ、研究データの国外移転、学習利用、サーバの場所等について、国として考え方を示し、適切な管理・利活用の徹底を図るといったことも記載されております。また、(3)ですけれども、AI for Scienceを支える次世代情報基盤の構築のところ、我が国の誇る研究データの管理・利活用のための研究データ基盤(NII RDC)、流通基盤SINET、「富岳」等のスーパーコンピュータ群を活用するとともに、HPCI(革新的ハイパフォーマンス・コンピューティング・インフラ)を中心としたユーザビリティの高い共用計算資源の戦略的な増強を行うといったことも記載がなされております。

少し飛ばさせていただきます、次のページ、13ページ目を御覧ください。こちら5ポツ目、研究施

設・設備、研究資金等の改革のところにも関連記載がございます。

(1) 先端研究設備等の整備・共用・高度化の推進のところですが、全国の研究者の研究設備等へのアクセスを確保する。このような研究大学は、SINETのセキュアで大容量のネットワークで接続することとする。これにより、先端機器群のスムーズな遠隔利用が可能となり、全国の研究者の機器へのアクセスを格段によくする。さらに、全国の先端研究機器群から生じるデータを集約することが可能となることから、これを体系的に保存し、幅広く研究者等の利用に供するといったことが書き込まれております。

また、下のほうですけれども、第4章、科学技術と国家安全保障との有機的連携の3ボツ、研究セキュリティの強化等のところにも関連記載がございまして、取り扱う研究データ等の機微性に応じた適切なアクセス制御を含め、情報システムに係るサイバーセキュリティの強化に向けて組織的に取り組むことも重要であるといったことの記載がございまして。

駆け足ですけれども、次のページに進んでいただきまして、こちらが、本ワーキンググループの第2回、第3回でも紹介をしまいたしましたが、AI for Science推進に向けた基本的な戦略方針でございます。3月31日に文部科学省決定がなされておりますので、御紹介をさせていただきます。

本戦略方針は、先ほど御紹介させていただきましたが、第7期科学技術・イノベーション基本計画の期間に当たる今後の5年間における計画となっております。

まず、上段のほうに背景ですとか、日本の強み、日本の課題を書かせていただいておりますけれども、その真ん中あたり、日本の強みのところに、本ワーキンググループの関連でございますSINET、NII RDCの記載がございまして、こういった情報基盤を有していることが日本の強みとして記載がなされております。

目指す姿としましては、AI for Scienceによる科学の再興として、AIが研究の自然な一部となる環境の実現、分野横断的人材が学術・産業双方で活躍、自律性と信頼性を備えたAI for Science先進国の地位を確立することとしております。

もう少し細かいところになりますけれども、16ページ目、次のページを御覧いただきまして、こちらが戦略方針の概要のスライドでございます。

先ほど申し上げましたが、今後5年間の計画となっております、集中改革期間と位置づけており、具体的な20のアクションを設定して、大胆な投資によりスピード感を持って取組を加速することとしております。

また、日本の強みを生かして、戦略的な国際連携による世界を先導する研究開発、新たなチャレンジとAI for Scienceの波及・振興、これを支える次世代研究基盤の構築、AIを高度に利活用できる研究人材の育成等を、関係省庁等と連携して強力に推進することとしております。それによって、研究環境と科学研究プロセスの革新により、自律性と信頼性を備えた研究国家としてAI for Science先進国の地位確立を目指すこととしております。

真ん中あたり、目的のところですが、科学研究の革新と科学的発見の加速・質の改革、研究力の抜本的強化と科学の再興、国際的優位性・戦略的自律性の確保という3本柱を掲げておまして、中期的な取組目標としましては、科学基盤モデル/エージェントやAI駆動ラボの活用により重要技術領域の先端的成果創出及び研究開発期間を10分の1にすることなどを挙げております。

こちらの下のところ具体的なアクションになりますが、一番最後の丸3、研究データのところが、本ワーキンググループでも関連するところがございます。戦略的価値の高いデータセットの特定と構築、自動化・自律化した研究設備等の整備と研究データ創出プロセスの標準化、AI時代に即した次世代情報基盤の構築・活用、データの一体的運用などの記載がございまして。

続きまして、次のページ、17ページ目を御覧ください。こちら前回は御覧になっているかと思っておりますけれども、具体的なアクションが設定されているというスライドでございます。

こちらの赤枠の中が本ワーキンググループでも関連のところになっておまして、丸3番の研究データ基盤システムNII RDCを2030年度までに容量5倍、AI化、また、丸4番、学術情報ネットワークSINETを2028年度までに2倍の高速化といったことの記載がございまして。

この後のページですが、18ページ目以降が、AI for Science推進に向けた基本的な戦略方針の関連部分の文章の抜粋でございますので、こちらは適宜お目通しいただければと思います。

駆け足になりましたが、事務局からは以上でございます。

【尾上主査】ありがとうございました。

最初のほうでは、これまでの議論、先生方から様々な御意見いただいたところ、委員からの意見のポイントの太字で書いているところで整理する形でまとめていただいているところでございます。また、後半については、政府関連の文書への記載状況というところでございます。

これらにつきまして、何か御質問等ございますでしょうか。よろしいですか。結構かと思しますので、早速議論に入りたいと思います。ただいまの文部科学省からの御説明も踏まえていただき、NII、黒橋所長より知識基盤の構築に関して30分程度御発表いただきまして、まだ時間はございますので、45分、50分ぐらい議論の時間を設けられればと思います。

それでは、黒橋所長、よろしくお願いいたします。

【黒橋所長】ありがとうございます。御紹介いただきましたNII、黒橋でございます。知識基盤の構築ということで、少しお話しさせていただきます。

まず、もうこれはここにおられる皆様に言うまでもないことでありますけれども、AIによる新発見が次々と行われております。医学、生命科学においても、長年の数学の未解決問題においても、宇宙の謎、理論物理学の新しい定式化、それから考古学も含めて、人文・社会科学も含めた大きな発見がずっとこの1年、特にこの半年ぐらい続いておりまして、これは加速度的に進んでいくものと、非常に広い意味でのAI for Scienceの成果が既に生まれ始めていると思います。

この資料は毎回のNIIからの御説明で使わせていただいているものでございますが、学術研究プラットフォームというものが少し右上の大きな箱でございますけれども、実験設備につながってデータを自動的に収集・解析、メタデータを付与して、これを管理・共有もするというのを御説明してきておりますけれども、本日は、この上に載っております知識基盤、一番肝になるといいますか、AI機能を含めた部分について御紹介をさせていただきたいと思っております。

私のお話は二部構成でございます。まず、前半NIIで、これは文科省からの非常に大きな支援をいただいて進めております言語モデルのチームサイエンスとしての開発、生成AI研究開発について御紹介させていただきます。後半、もう少し研究者のAI利用シナリオみたいなものを整理して、それに対してどういう機能を提供し、そのためにはどれほどの計算資源が必要かということ、具体的な数字がやはりまず議論のスタートしてある必要があると思しましたので、そういうお話をさせていただきます。

LLM - jp / LLMCということですが、これは先ほどと繰り返しになりますけれども、本当にこの1年ほどでLLM、それから、それをエージェントとして動かすAIエージェントについても大きな進展があり、Deep Research機能というのは、もう本当に日々研究者が使うものになってきていると思っておりますし、AIが著者に入るか査読をするか、この辺りも非常にコントラバーチャルですが、そういうことの実験的会議も開かれています。近々でも、ClaudeのCoworkですとか、そういうものがソフトウェア業界に大きな衝撃も与えていますし、4月になって出たといいますが、存在が分かったMythosというのは、強過ぎて一般公開が見送られると、過去何十年間分のソフトウェアバグを発見しているという世界になってきております。

こういう中で、やはりLLMをきちっとつくれる力というのは日本においても必要ではないか、あるいは、それをきちっとオープンにすることですとか、日本についても理解を深める、そういう趣旨でLLM - jpというものを3年ほど前に立ち上げまして、2年前から文科省の支援をいただいて、LLMのセンターをつくっております。最初はいろいろ試行錯誤がございましたけれども、LLM - jp - 3というシリーズ、4というシリーズで、この4月に入りましていろいろ新しいモデルを公開できるということまで来ております。

この4月3日に公開いたしましたのがLLM - jp - 4という言語モデルでございます。右側にございますように、8ビリオンのDense、それから、32ビリオンフルパラメータ、アクティブ3ビリオンというMoEモデルを公開しております。学習トークン量が前のモデルに比べますと5倍ぐらい大きくて、十数兆というスケールでございます。パラメーターサイズについて言いますと、Qwenの大きいものが300ビリオンというようなサイズ感、それから、いわゆるフロンティアモデルのChatGPTですとか、そういうものが数兆、先ほど申しましたMythosというのは10兆ぐらいと言われておりまして、ビッグテックは3桁以上大きいGPU資源があるわけですが、その中でQwenオープンウェイトモデルと並ぶぐらいのものができてきたということでございます。

もうごく簡単に学習レシピということですが、これも世界でどんどん進んでいる状況をウォッチしながら進めていますが、まずフェーズ1として、一般的なウェブのコーパスが大きいわけですが、広範な知識を獲得すると。

その次に、フェーズ2として、数学ですとかコーディングの基礎能力を伸ばす、そういうデータの割合を高めた学習をします。この辺りではデータのクオリティを上げる、それから、どういうデータをどういう割合で学習すれば性能が上がるか、そういうアブレーションテストもしっかりとしながら進めていると。

次に、中間学習というフェーズですが、この辺りでは、QA能力といいますが、指示追従性を

伸ばすような学習をまず中間学習の最初のフェーズで行いまして、フェーズ2では、これをロングコンテキスト化すると。今のモデルは64K、6万4,000トークン、約6万4,000単語のデータが扱えるというものになっておりまして、これ全体で約12兆トークンの学習をしたものとなっています。

このモデルの総合的なパフォーマンスが、ここでお示ししておりますグラフの一番右側ですけれども、これ評価データセット自身も、LLM - jpとしてエバリュエーション、ベンチマークをつくっております。このバージョン2というのは、最近の直近までの非常に難問も含めた問題、ベンチマークになっています。このクオリティが、一番右、順番に青いものがQwen3の8ビリオン、その隣がgpt - ossの20ビリオン、緑のものが以前まで公開しておりましたモデル3ということですが、一番右の2つが今回公開したもので、Denseの8ビリオンとMoEの32ビリオンということで、見ていただきますとおりですけれども、Qwen3ですとかgpt - ossとほぼ同等のところまで来ております。ここまで来ますと、皆様が多分ChatGPT等で使い慣れておられるようなことが、そのクオリティはもちろいろいろございますけれども、ほぼできると。ここにありますけれども、OpenWebUIというオープンソースのインタフェースが公開されておりますので、これは広く使われておりますけれども、そこに我々のシステムを載せますと、例えば、添付資料をつけることもできますし、そこで、オープンなLLMは、「フロンティア」LLMに比べてどんな長所がありますかというような質問をしますと、ウェブ検索もこのインタフェースにある種組み込まれていまして、オープンなウェブ検索ツールをたたいて、その結果も含めて、右下全体がそうですけれども、こういう回答をすることができます。テーブル形式で結果をまとめて、それから、このデータをCSVファイルで吐くとか、そういうことが全体的にできます。

これができるようになったのは、やはり先ほどのようなレシピ、学習の方法で指示追従性を上げているということが大きくて、質問に対して、それを適切なWebQueryにして、その結果を持ってきて、それを解釈して、もう一度プロンプトに入れて動くと、こんなことができるレベルになってきています。

それから、今、本当に研究者のふだん使いになってきていると思いますけれども、Deep Researchということもできております。ユーザの入力に対して、そのモードを判定して、何を調査したらいいかということ立案し、クエリの候補を出して、クエリの結果を得て、それをベースに鑑定を拡張する。このループを何度か回して、最終的な候補を決めて、分かったことを統合してレポートを書く。まさにDeep Researchが、今様々なサービスがありますけれども、行っていることができまして、例えば、これはLLM - jp - 4のモデルで、Niiの3基盤の一つですけれども、CiNiiにある論文ですとか研究者のデータをベースに、それをRAG的に活用して、例えば、防災の研究者で既に情報分野の研究者と共同研究を行っている人を教えてくださいと言いますと、ちょっと小さくてあれですけれども、先ほどのような流れで立案して、データを取ってきて、知識統合して、こんな先生がこういう研究をしていますよということをもとめて示すということができるレベルになっております。

この公開データに対して、4月の頭に公開しましたけれども、産業界でも非常にポジティブに受け入れられていると思っております。まず、クオリティが一定程度であるということ、それから、Apache2.0ということで、派生モデルの展開等にも高い自由度がありますし、日本としてこういうモデルを持ってオンプレで動かせる、ローカルに持ってこれるということでのデータ主権、いろいろ機微なデータを使いやすいというようなことがございます。

左下では、既にさくらインターネットでこれを活用することも、ホスティングを検討されていますし、さらに、フルオープンで公開していますので、それに対する様々なモジュールの追加開発ですとか、いわゆるインフルエンサーの方々の歓迎コメントもございます。官公庁等で資料を検討しているという声もお聞きしております。

これはコメントの具体例ですので、タイトルぐらいをばらっと御覧いただければというぐらいのイメージでございます。

もう一つ重要なこととして、画像も扱えるビジョンランゲージモデル、これにつきましても、4月14日に公開しています。これは先ほどの8ビリオンをベースに、プラス1ビリオン、つまり10億パラメータ画像を扱うように追加したものでございまして、画像と言語のデータを3,000万件という規模で整備をして、ビジョンランゲージモデルにしたものです。

例えばですけれども、この左側の例ですと、これは何ですかというと、文字として、北の丸公園というのが大きく御覧いただけだと思いますが、その隣に、ちゃんと東京大学入学式ということで、イベントとしてはこちらを出す。あるいは、この右のアイコンのようなものも解釈して、こういう応答もできます。

これはいわゆる自然画像についての応答ですけれども、この機能は非常に重要でございまして、例え

ば、文書画像といいますが、論文の図表などを解釈して動くためには、これが一番基本的なものになりますので、それで、そういう機能も既に一定程度備えていますけれども、その辺りも高めていきたいと。

実はこれも、ビジーなスライドですけれども、棒グラフの一番右を見ていただきますと、Qwen3のビジョンランゲージ8ビリオンと同等の精度を達成しておりまして、その右はgpt5点幾らだったと思います。それは強いんですけれども、十分オープンモデルとしては強いレベルになってきているということでございます。

これが前半のまとめのスライドですので、これだけ少し丁寧に述べさせていただきますけれども、NIIといいますが、これは、日本中の研究者の方に参加していただいて進めているプロジェクトでございます。このプロジェクトによって広く活用されているオープンウェイトモデルQwen3と同等レベルのソプリンのLLM、VLMを開発することができまして、これはNIIとしても、あるいはLLMCとしても、この社会実装に向けていろいろ取組を始めています。

Llamaというものがかなりオープンモデルとして広く使われましたけれども、その勢いはもうかなり衰えていますし、今はQwen、中国のモデルが活用されていますけれども、2026年3月にはその主要研究者の退職が報じられるなど、企業主導のオープンウェイトモデルの開発というのは、どういう見通しか不透明な部分がございます。

それから、クローズドなモデル、いわゆるフロンティアモデル、ChatGPTですとか、Claudeですとか、そういうものも著しい発展を遂げておりますけれども、やはり収益上重要なドメインですとか言語に集中しております。AIと人類、本当にこれから大きな問題といいますが、共存の仕方を考えていれないといけないわけですし、その中では人文・社会科学も極めて重要と思われませんが、そこまで十分にカバーされるかどうかは何とも言えない状況と思えます。

それから、クローズドモデルはもちろんですけれども、オープンウェイトであっても、学習データとか開発レシピは公開されておられませんので、そういう意味では、いわゆるScience for AIの観点からは、あまり知見は得られないということが出来ます。

完全なオープンモデルの開発としては、ワシントンのAllen InstituteがOLMoという、これは大変すばらしいプロジェクトですけれども、ここでは英語に限定したオープンなモデルの開発が行われています。これはアメリカにおいて、あるいは、そういう立場はもちろんあっていいと思うんですけれども、一方で、多くの国は、いわゆる共通的な英語の世界と自国の言語・文化というものを両方きちっとカバーできるモデルを持ちたいということで、我々の活動に対して、既に韓国、インド、タイ、フランス、ブラジルなど、連携を始めております。

ソプリンAIと言いますと、一国で完結するクローズな雰囲気は響かなくもないかもしれませんが、それは全く逆でございます、逆に国際的な協力の下で推進するということが非常に重要だと考えております。

ちょっと長くなりましたけれども、そういう基本的な技術力といいますが、しっかりしたモデルはあるということをベースに、これからNIIとしてといいますが、考えていくべき知識基盤のAI機能について少し整理したので、御紹介させていただきたいと思えます。

AI for ScienceにおけるAI利用シナリオというものをまず整理して、かつ、それに必要となる計算環境というものも考えたと。ここで整理したいAI利用シナリオは、すなわち、今回の知識基盤が機能提供しようと考えているものでございますけれどもAI for Scienceですから非常に広い分野がございますので、これは今後いろんな分野の方々と相談をしながら深めていくものですが、その出発点の整理であります。

これも繰り返すまでもありませんけれども、学術研究のあらゆる側面がAI化していくということで、どのようにAIが利用されるかということ一度整理して、そのために必要となるGPUの計算資源がどの程度であるか、NIIとしてどう考えていくかということ整理したものであります。

一応誤解のないように、ここは本当に強調したいと思うんですけれども、こういうサービスをNIIがする場合に、先ほどの我々のほうでというか、日本として開発したLLM - jpとか、そういうモデルだけを動かすのではなくて、その時点での良いオープンモデルを動かす、幾つかのものを動かすということを考えております。

ただ、この資料では、そこでどれぐらいのGPU計算環境が必要かという試算においては、LLM - jpのモデルを使っているということでございます。もちろん、このモデルを使う意味もいろいろあると思っておりますけれども。

まず、現在の実態みたいなものを少し整理したのがこのスライドでありまして、現在、一般的な利用はやはりフロンティアモデル、このクローズドモデルを、無料または有料版を使っている。それが

ら、当然研究者の様々なデータですとか論文ですとかRAGとして使用したり、これにおいては NotebookLMが一般的かと思えます。自分自身のデータでLLMを追加学習したいということもあるかもしれませんが、そこまでされている方はあまりないですし、クローズドモデルの場合にはそれは非常に難しい。

それから、秘匿性の高いデータ、研究の過程のデータはそもそも一定程度の秘匿性がありますけれども、さらに医療ですとか様々なデータがあります。これについては、入力データは学習に用いないという契約の下で、あるレベルまでは使うですとか、それ以上は心配なので使わないという状況かと思えます。

こういう状況というのは、特に無料版というのは機能が制限されていますし、有料版というのは非常に組織単位で見れば大きなコストです。それから、秘匿性の高いデータは必ずしも活用できない。それから、RAGですとか追加学習の自由度は限定的でありますし、何よりも重要なことのひとつとして、実験の再現性の確保というのがあると思えます。フロンティアモデルというのは、どんどん変わっていった新しいバージョンになるわけですから、ある時点で研究のためにいろいろ処理したことを後から確認できない。それはパラメータを変えて試すとか、そういうことはもうできなくなっていく。そういう再現性のなさというのは非常に大きなことで、これはクローズドモデルに完全依存することの問題のひとつと思えます。

一方で、オープンモデルの活用ということも徐々に進んできているわけでございまして、オープンモデルを手元に持ってきて動かすということです。その場合には、RAGですとか追加学習の自由度は一定程度ありますし、秘匿データも扱うことはできる。ただし、これを行うためにはGPU環境を用意する必要があるのと、やはりモデルをそういうふうに加えて追加学習するですとか運用する、そういうノウハウが必要で、ここまで実践できる研究者は一定程度限られてしまうというのが現状かと思えます。

それから、選択できるモデルとしてQwen等がありますけれども、これはもう既に指摘したことですけれども、それがどれくらい継続されるか不透明であるということも言えますし、この点はまたこの後で議論させていただければと思えます。

こういう中で、NIIとしては、学术界全体へオープンなモデルのAI利用環境を提供していく責務があるのではないかということで、そのための環境について検討しておりますし、最初はまずそのための利用シナリオを整理して、計算資源を試算したということでございます。

AIの利用シナリオですけれども、2タイプ、大きく「バッチ型」と「対話的サーバ型」のサービスに大別できると思えます。

バッチ型の処理につきましては、ここに挙げました4つぐらいの使い道があるだろうと。

まず、やはりRAGをつくるということは各研究者にとって非常に重要なことで、そのためには論文ですとか実験ノート等をOCR等でまず読み込み、それから、それをベクトル化し、インデックス化するということで、これで研究者あるいは研究グループの知識ベースを構築するということになりま

す。

2番目として、AI-readyデータの整備ということで、実験のノートですとかデータを対象に、クレンジング、構造化、メタデータ付与、いろいろ統合等するという処理があります。

3つ目として、LLMの追加学習を行って特化モデルを構築する。オープンな学習データももちろん大量にありますし、それから、そこに自分のデータを加えて、自分のといたしますが、その分野の特化したモデルをつくり、これによって解析精度を向上させたいということはもう非常に大きなニーズだと思えます。

4つ目は、一般モデルでいけるかもしれませんが、あるいは、特化モデルを学習した上で、一定規模の実験データに対して解析推論を行う。これはまさにAI for Scienceのメインの活動ですけれども、そのためのバッチの推論ですとか解析の処理を行います。これは一定程度のデータが取れば、それを解析する、そういうサイクルになると思われま

す。

2つ目が、対話的なサーバ的な処理で、これはAI LLM特化モデル等がサーバとして常駐していて、それに対話的に利用するという、ChatGPTのインタフェースのようなイメージを持っていただけたらと思えますけれども、まずはアイデア出しですとか、研究計画ですとか、共同研究者の検討などを行う。先ほどちょっと見ていただきました防災の研究者で情報研究者と一緒にできる人を探すとか、そんなようなイメージでありまして、一つ重要なこととして御指摘させていただきたいのは、やはり社会課題解決に向けて研究していこうと思うと、必然的に分野横断的なものになり、その研究テーマの探索ですとか、あるいは、共同研究者の探索というのは非常に重要でかつ難しい課題でありまして、ここはAIの活用が非常に有効であろうと。

2つ目は、日常の研究活動の支援ということで、実験ログを少し取りまとめたり、実験ノートを構造化するという日常的な活動もあるでしょう。

3つ目は、報告書・論文の執筆支援ということで、一定程度のものを、中間段階のレポートもあり得ますけれども、下書きを書いたり、構成案を考えたり、翻訳したり、あるいは、文書として構成する、そういう機能がございます。

4つ目として、論文・研究提案書などの査読の支援。これも現在は、フロンティアモデルの場合には、クローズドモデルの場合には、そこに研究論文、発表前の論文ですとか採択前の研究課題提案書というのは本当に秘匿性の高いものですので、これをフロンティアモデルに入れるというのはなかなか難しいという中で、きちっとしたサービスがあれば、それは十分活用されると思います。実際、いろいろ学会等でも研究者の方と話していると、もう本当に査読の負担が大変だと。AIに入れたらいろんなことが分かるのに、今、全然それができない、何とかしてくれということは、よく私もコメントを伺います。

それと、重要なのは、論文が査読できる機能というのは、実は論文をブラッシュアップできるというんですか、研究を自分自身でブラッシュアップできる機能ですので、今まで国際会議に出して、いろいろコメントが来て、それでアップデートしていくわけですけれども、そういうのが実は手で非常に早いサイクルで回せる。そういうメリット、そういう使い方もございます。

今申し上げましたようなことに対して、いわゆるテキストベースの言語モデルもございませぬけれども、先ほどのRAGでもそうでしたけれども、NIIがこれまで蓄積してきた学術界の様々なデータ、論文ですとか研究者に関するデータ、過去の共同研究のデータですとか、そういうものを構造化して知識グラフとしてまとめて、これをRAGで活用する、基本的なものと活用する、これも重要な機能であると考えております。

今のようなシナリオで、ではどれぐらい計算資源が必要かということを試算いたしました。ここではLLM - jp - 4の8ビリオンを一応試算に使っております。パラメータ数8ビリオンということですので、物理サイズとしては17ギガバイトぐらいあれば動くということで、比較的この場合軽量です。それから、追加学習とかスループットですけれども、これは我々の構築時の実験等での実際の値ですけれども、4ビリオントークンぐらいの学習が、H200、8GPUを想定した場合に、1ノード1日程度でできます。4ビリオンというのは40億トークンですので、それなりの一定の規模ですので、多くの分野の追加学習というのは、一つの目安として、この程度か、これの数倍、そうすると数日程度ではできるということになります。

推論環境はA6000等でもできますし、H200ぐらいのものがあれば、1GPUで複数のモデルを動かしたり、vLLMで非同期・マルチセッション型のサービスをするということも可能になります。

これで先ほどの活動を試算するわけですけれども、研究者がどれぐらい使うかというのは、もうこれはどんどん増えていきますので、なかなか試算が難しいのですが、逆に一回試算しないと議論ができないということで、ここではもう本当にたたき台として試算をさせていただきました。

まず、先ほどのバッチ型のRAG化とかAI - readyデータを作る、これはどれぐらいかという試算ですけれども、そのRAG化につきましては、研究者が80万人と、ここでは試算しております。50万人ぐらいの研究者プラス研究補助者、それから、大学院生が一定程度というイメージですけれども、こういう人が年間2,000文書ぐらいについてのRAGの構築作業をする。その作業で一番重いのは、やはり論文をLLMでOCR化するところございまして、図表等も含めて、この辺りがマジョリティになりますけれども、年間12ノードぐらい、これは1ノード8GPUが載っているH200というイメージですけれども、これぐらいの規模になる。

それから、AI - readyのデータを作るのが、5,000研究グループ、そこに属する5万人ぐらいの方が実験などをされるから、年間100回ぐらい、3日に1回ぐらい、どんどんデータを整備していくというイメージでいきますと、これが3.6ノードという規模になります。

それから、追加学習は一定程度重いわけですけれども、ここまでやるグループは先ほどよりは少し少ないかもしれませんが、つまり、特化モデルをつくるということですね。これは1,000研究グループが毎月やったとして、4ビリオントークンぐらいの追加学習をやると、これが32.9ノード。

それから、実験データ等を実際に解析するというのは、これは5,000グループぐらいが年に20回ぐらいやるとしますと、22ノードという試算になります。これはH200を1ノード2時間回すぐらいの計算になるのではないかと想定です。

2つ目の対話的なサーバ型の利用です。これは、まず研究者の見積り、ちょっと前後しましたけれども、研究者が、URAとか研究支援者を含めて、大学・研究機関に所属する方が50万人ぐらい、大学院生が28万人、学部、1回生から3回生も使うかもしれませんが、ここでは4回生として65万人と。こ

の人たちの対話は平均これぐらいの回数として想定しておりまして、この研究者1日30回という、その1回は、PDFを添付するようなこともあるでしょうし、3往復ぐらいして壁打ちをしているというようなイメージで入出力のトークン数を見積もっています。

右側が、我々のモデルも含めて、実測値で試算したものですけれども、大体3万トークン/ノード秒と。1ノードあれば、1秒に3万トークンぐらいさばけるだろう。この試算を基に、この左ぐらいの対話をさばくためには、127ノードが必要ということになります。

今のものを全体的にまとめますと、この右側にごさいますように、約200ノードぐらいあれば先ほどの見積りには対応できるだろうと。ただし、これは単純に足しましたので、適切なといいますが、妥当なレスポンスで対応するためには、少なくとも倍、400ノードぐらいのものが必要だと。

もうこれはチップもどんどんよくなりますけれども、今これをH200、1ノード、8GPUぐらいの環境で想定しますと、NIIのLLMセンターの我々のこれまでの経験ですと、民間クラウドで調達すると、この規模というのは年間140億ぐらいになります。

それから、オンプレでこれを整備した場合はどうなるかということ、これはmdxの関係者の皆様に試算していただいたところ、システム構築が300億、保守・運用が年45億、これを5年間活用するとすると、平均は105億ということで、民間クラウドよりは当然経済的ではありません。

ちょっと下に書かせていただきましたが、その経済的観点もごさいますし、加えて、大学での計算基盤運用力の維持・発展ですとか、そういう人材育成の観点からも、オンプレ整備には大きな意味があると思っております、その場合の利用者負担の考え方も含めて、mdxの皆様と前向きな議論を開始したところでごさいます。

まとめといたしまして、AI活用を日本の学术界全体に広げるために、もう本当に大学院生も含めて、AIを使った実験試行が簡単にできるような環境を提供するためにということで、やはりオープンモデルで共通AIの利用環境をこのような形で提供する。そこではバッチ型の計算ですとか、対話的なサーバ型の利用ができると。それによって追加学習、ドメイン特化学習のモデル構築等が一定程度できるということと、やはりそこは非常にノウハウもあるところですので、そのノウハウの支援ということも重要であると思っております。

追加学習は当然そうですけども、加えて、プロンプトをきちっと設計して、AIエージェント的にいろんなものを使いこなす、これについても非常に様々なノウハウがありますので、その意味でのコンサルテーションですとか、あるいは、NIIがコンサルするというよりは、情報共有の場があって、そこで皆さんが意見を交換してノウハウを共有する、そういうことも必要と思っております、今日御紹介したようなサービスをするためには、400ノードぐらい、オンプレで年105億と、こういう規模感でありますので、議論の出発点とお考えいただきまして、議論させていただければと思います。

以上です。ありがとうございました。

【尾上主査】黒橋先生、ありがとうございました。

それでは、ただいまの御説明に関しまして御質問、コメント、御意見等ございましたら、挙手にてお知らせいただければと思います。いかがでしょうか。

宮田委員、どうぞ。

【宮田委員】いろいろ必要で実際使っている機能があり、これができたらすばらしいなと思いがら、お話を聞いていたんですけども。

前半のところの最後だったか、ソブリンAIをつくるのに国際的に協力していくというお話を、むしろそうしていくべきだというようなお話をされていたんですけど、そのロジックを分かりやすく説明していただきかけたのと、それをやろうと思ったときに、何があつたらできるのかということですかね。お金があれば別にこのコラボレーションはうまくいきますという話なのか、もう少し別のサポート、制度的なものとか、国際的な話なので、輸出管理的な話もちろんあると思えますし、そういったところのコラボレーションを阻害する要因が例えばあつたりするのかなど、難しさと、なぜそもそも国際的に進めたらいいのかということを教えていただきたいなと思いました。

よろしく願います。

【黒橋所長】ありがとうございます。

まず、なぜ必要かですけども、我々のところは完全オープンでやっておりますので、そのノウハウは外国の方も活用、見ることは十分できます。が、やはり実際の学習のノウハウの最後のところというのはもうちょっと、結局のところ、しっかり話さないと分からないというところはありますので、我々も中国の方に何かを聞くとか、インドの方とそういうディスカッションするという、もう一歩踏み込んだディスカッションは当然必要になると思います。

複数の言語をどんなバランスで学習したときにどういうことが起こるか、かなり規模が大きいので、

実際我々も小さいモデルでアブレーションとかはしていますけれども、それを大きなモデルで何回もできないわけですね。そうすると、そこをしっかりと役割を分担して、こういう学習の仕方ではこういう結果だったということがディスカッションできることは重要だと。

それから、言語が違っても、例えば、やっぱり分かってきていることは、数学とかコードのデータは非常に重要で、これは一定程度英語で開発されていてもよいわけです。じゃ、数学的な学習データが十分にあるかということ、そういうわけでもなくて、我々、Open Source AI Definitionという、オープンソースできちっと使えるAIのモデルという規範に準拠する形で考えているんですけども、その規範で使えるデータというのは本当にそれほど多くありません。そういうことで、今、例えば、ブラジルなんかと議論しているのは、そういう数学のデータの構築を一緒にやって、共にオープンなものをつくっていかうと、そういうこともございます。

ですので、実開発の様々なノウハウをもう一步踏み込んでシェアするということと、やはり共通で使えるデータセット等もございますので、その辺りは国際協力していけばいいのではないかと考えております。

何が必要かは、一つは、やはり話し相手となってもらえるための実力といたしますか、基盤的な力が必要で、今我々はそれなりの立場に立てていると思うんですけども、そういう研究開発力を持つことが何よりも重要で、それがあれば、あとは、放っておいてもといたしますか、国際協力が活発になり、また、ハブとなっていくこともできるかもしれませんし、それは非常に重要なことだと思っております。

それで、我々、輸出制限とか、そういうレベルのところにはまだ全然至っていないですけども、逆に、LLMの我々の活動としては、日本のAISI、AIセーフティ・インスティテュートとの協力は十分やらせていただいている、技術的に交流というか支援をさせていただいていますし、もちろん産総研さんとも協力させていただいていますし、その辺りも本当にイギリスなんか強いんですけども、日本としてもしっかりした活動をして、それこそこれは世界共通の問題ですので、その辺りの議論も含めてちゃんとやっていくと。

その辺りも基本的にはソブリンAIというか、オープンな考え方での協調だと思いますので、そこにAIセーフティの課題への検討において、LLMといたしますか、生成AIを作っている経験値というのは非常に大きなバックグラウンドになると思っています。

【宮田委員】ありがとうございます。

【尾上主査】それでは、若目田委員、どうぞ。

【若目田委員】ありがとうございます。

何かめっちゃくちゃ基本的なことになってしまうんですけども、これだけかかるというコストのところは分かったんですけども、これを入れる効果みたいなどころというのは何か定量的にあるのかどうかということも伺いたかったです。これは無料で使えるので、どんどんみんな無料で使いましょうというものなのか、国費を入れているので、やはりそれに相当する回収のコミットメントが必要なものなのか、その辺のところはそもそも分からなかったの、教えていただきたい。使うたびに、基本は、例えば、今まで研究でそれにコストがかかっていた、手間がかかっていたのがかからなくなったので、それは使う人もメリットなので費用を負担するのかとか、そういう基本的な考え方のところを伺いたかったのが1つです。

2つ目は、やっぱり秘密データというか、オープンにしたいくないデータにすごい価値がある、例えば、企業が抱えているデータというのは、これから多分非常に価値を持つので、むしろそれをどうやって社会に使うかって多分問われていると思うんですけども、研究もそうかもしれません。なので、そういうような部分をいわゆるクローズのベンダーのLLMとかに入れてしまうのは、これは多分我々も感じている課題なので、それにこのオープンのもは対応できるというのは大変重要なメリットかなと思ったんですけども。これがオープンでということは、多分、仮にこれが企業が使うんだったら、各企業とかがクローズで閉じた環境で使うので、このクローズのデータが担保されるという意味なんではなかうかね。なぜオープンはこのシステムだとデータの秘匿性が保てるのかみたいな、その辺のところの仕組みを伺いたかったということが2点目です。

以上でございます。

【黒橋所長】ありがとうございます。

まずコストについてですけども、そこはもうちょっと深い試算をしていかないといけないと思っておりますが、今、もうAI for Scienceで、本当に誰でもが使うということになると思いますが、その人たちが、例えば、ChatGPTの安いほうの有料プランを使うということになれば、これはもうはるかに、500億とか、そういう桁になります。

無料プランを使っていればいいのかというのは、最初に指摘させていただいたいろんな問題があるので、一定程度そういうことを使っていく中で、これも一つのオプションになるということ。それから、この105億を運用することによってこういうサービスができるんですけども、そのときのいわゆる受益者負担の考え方もこれから整理していく必要があると思います。

ただ、本当に研究をぱっと試してみたいというときに全部課金するのかという、それは逆に研究のいろんなアイデアとかアジャイルな開発というかが難しくなりますので、一定程度のところまでは日本のアカデミアであれば無料で活用できて、少しヘビーに動かす場合にはそれが課金されると。これは今後mdxの方々も含めて、日本の情報基盤の皆さんといろいろ検討していきたいところですけども。とにかく日本のアカデミアがぱっとそういう支援が使えるという環境をいかに提供できるかという問題かと思っています。

それから、2点目のセキュアな環境ということですけども、もちろん、ここでログですとか、そういうものの使い方についていろんなターゲットをきちっと設計をして、企業の方とどういう関係していくかということはこれからしっかり議論しないといけないところですけども、契約があるからそれで安心してということとは別の選択肢として、ここは非常にセキュアな環境でデータは安全だと、国内で運用していると、そういう環境が提供されることは十分意味があると思っています。

一方で、そうでないもうちょっと緩やかなログについては、逆に検索エンジンの時代の議論と同じだと思うんですけども、Googleとかがどうして強いかというと、検索ログを徹底的に集めてそれを活用できるからという面もあるので、そこまで秘匿性の高くない、しかし一般的な研究活動においてということが議論されるかということは、ユーザが許す範囲できちっと国の中で把握をして、その結果をまた国の次の政策に生かしていくとか、そういうメタな活用方法は一方であるかもしれませんが、いずれにしても、そういうコントロールが秘匿性に依拠できるということは、こういう基盤を国の中できちっとサービスしていくことの大きな意味かなと思います。

【若目田委員】2点目は、そういう意味では、そういう私企業が運営しているところのリスクというものがないんですよ。契約なり約束事で、NIIさんというか、この運用の中で、例えば、クローズドなことが担保できるとか、そういうコントロール可能という運用の仕方ですね。

【黒橋所長】はい。

【若目田委員】分かりました。

1点目のところは、むしろお金を取ってくださいという意味よりは、各研究者がこれだけ提供されたら、今まで何らかの過程でできなかったこと、当然その研究の成果に関しても、むしろそっち側のほうを上げていくというか、そういうところのコミットメントで、むしろ推進の起爆剤になるような、やれない理由とかやらない理由みたいなものがこれでなくなるというぐらいのポジティブなものに移ったほうがいいんじゃないかなと、そういう意味も含んで申し上げました。

【黒橋所長】ありがとうございます。

そこは本当にそうだと思っておりまして、これがまず試せるという環境があることの大きさだと思っています。

ちょっと聞いた話ですが、東大の先生が本郷で飲み屋さんに行ったら、後ろで学生さんが話していて、おまえ、その課題をやろうと思ったら20ドルじゃ駄目だから200ドル課金しないといけないよみたいな話をしていたそうなんです。だから、今、結局それはどういうAIが使えるかということで教育とか研究の質が変わっているようなところなので、その基本レベルはこれできちっと提供したいというふうに思っています。

【若目田委員】多分、今みたいなものをもっと定量的に測っていけば、かなり重要なあれだと思っています。

【黒橋所長】そこが次は重要なところなんですけれども、エピソードトークしてもあまり意味はないと思うんですが、ただ、これもAIの技術レベルについても、それから、研究者の活用についても、もう加速度的変化なので、ある瞬間でアンケートを取って、あなた、どれぐらい使いたいですかと言ってもあんまり意味を持たない気もしております、そこは我々としても必要性は十分感じておりますけれども。

御指摘ありがとうございます。

【尾上主査】江村委員、どうぞ。

【江村主査代理】ありがとうございます。

今日お話しになったことに対して個別にはあんまりないんですけど、AI for Scienceという時代になってきたときに、NIIの今学術研究プラットフォームと言っているものが貢献する領域と言ったらいいのかな、それが、従来型のイメージで研究が進むという領域に限定されるイメージになっていな

いかという懸念を持ちます。だから、言語系とかはそれでいいと思うんですけど、今のAI for Scienceって、リアルな実験とかなんかも含めて、研究のスピードを上げたり、探索範囲を広げたりというところにすごく価値が出てくるわけですよね。そういう動きに対して、やっぱりこの学術研究プラットフォームというのをコアの部分だけ独立した形にすると、先生の資料にあるように、バッチ処理とか対話型とかという、そういうところには価値が出てきますという構造なんですけれども、前回からも山地先生とかにも申し上げたりしているんですけど、実験観測施設はこの学術研究プラットフォームの外側にあります、HPCはこの外側にありますという絵を描いていると、これからこのAI for Scienceでどんどん進んでいこうとしているところへの貢献が小さくなってしまわないか。論文を書いたり、そういうところのプロセスは改善されるというのはどの領域でも共通なんだけど、もう少しこの基盤が持つ価値を出していくという観点で言うと、もう一歩進んだ検討や表現があるのではないのかなというのが、これまでの回も含めて私ずっと思っていることで、このNIIさんが書かれている今出ている絵そのものをもうちょっと変えたものをつくりませんかというのがいつも提案していることなんですけど。

じゃないと、やっぱりLLMの価値を出すというのは、今日おっしゃったところで、そのとおりだと思うんですけど、そこで止まっちゃっているのがもったいなくないですかというのが一番聞きたいことなんですけど、いかがでしょうか。

【黒橋所長】ありがとうございます。

この絵はアップデートしないといけないとは思っているんですけども、今の御指摘に対して幾つか申し上げるとすれば、まず、mdxというのはこの外にまさにありますけれども、ここはもうかなり踏み込んだ議論といいますか、相談を始めさせていただいています。もうこれは計算資源と表裏一体ですので、どこにどういう計算資源を割り当てて、どんなふうにコントロールしてというのを含めて、それは計算資源の方と一緒にやらないともう絶対成り立たないことで、何を知識基盤あるいは学術研究プラットフォームという呼び方の問題はあにせよ、そこはもう対等なコラボレーションすると、そこはもう我々は踏み出しています。

それから、実験設備については、ロボット実験とか、そこまではNIIの範囲にぱっと入れるのは難しいですけども、これまでに御説明しているのではないかと思いますけれども、かなりそことシームレスにつないで、もう実験データがどんどん入ってきて、それをどんどん処理していくと、そういうデータも付与していく、そういうフローはつくっていきたくて思っていますので、そこももうどんどんそちらのほうに出ていきますし、今年度始まるいろんなプログラムで、NIIとそういう実際の実験施設を持っておられるところとのコラボレーションというのはもうかなり増えていきますし、それから、国の中でもSPReADですとかARISEとかというのはどんどん進んで、そこでこの基盤を使っていた上でのコラボレーションを考えていきたいと思っていますので、こちらに閉じ籠もるつもりはなくて、どんどん進んでいくと。

それから、今日の私の説明の範囲が言語言語していたかもしれませんが、AI for Scienceのモデルも、今のところはまたやはりベースはトランスフォーマーで、そこにどんなデータを入れていか、あるいは、ツールコーリングでどんなデータを呼ぶかということで、それは今日ちょっと申し上げた追加学習というのは、少しそういうイメージもあります。ですので、各ドメインでどんなデータの追加学習が必要かということは、この基盤も一定程度支えられると思いますし、繰り返しになりますけれども、SPReADとかARISEの中で、そういうAIモデルの各分野特化ということについても一緒に考えさせていただきたいと思っています。

【江村主査代理】何でもパーフェクトなものをやろうとかというのはないと思うんですけど、絵を描き直されるということで言ったときに、今、知識基盤と結ばれているところが、一旦このプラットフォームに入ってきた中で知識基盤と結ばれているんですけど、そのリアルなものがこの外側で結構起きたときに、この知識基盤というのがどういうリンクの仕方があるのかというのは、いろんな可能性がありそうな気がしていて具体的なイメージを出せないのがあれなんですけど、その辺に次のステップがあるんじゃないかなみたいな。

【黒橋所長】それは御指摘のとおりと思うんですけども、想像レベルではなくて、もう今年度本当に1,000件ぐらいの研究が進んでいきますので、そういう方にもぜひ一緒に考えましょうという窓口を開いて、それぞれのドメインでどういうことが行われて、研究者がどこでもっとこれがやりたいということがあるのか、それをしっかり相談させていただきたいと思っています。

【江村主査代理】よろしくお願ひします。

【千葉委員】すみません。今のことで質問してもよろしいですか。

【尾上主査】じゃ、千葉委員、先にどうぞ。

【千葉委員】ごめんなさい。江村委員がおっしゃったのは、実験装置が枠の外に出ているからあまりイメージが湧きにくいということなんじゃないかな。これ、実際は、実験装置というのは、知識基盤が実体を伴った何なのかはあまり分からないんですけど、今恐らくモデルが動いていると、それぞれの実験装置ですとか研究者のPCも含めてですけど、モデルとネットワークで直接つながっていて、直接何かやる、ロボティクスも多分そうやっていらっしゃる先生、大勢いらっしゃると思うので、それではまだ不足だということ。

【江村主査代理】今おっしゃったとおりで、これからのAI for Scienceってロボティクスを入れて、実験がどんどん自動で行われて、そこからどんどんデータが出てくるわけですね。出てきたデータを、どこかモデルがあったりなんかするかどうか分からないんですけど、解析をして、実験を人間が介さなくてもどんどん次のステップに回していくというようなのがAI for Scienceの新しい時代で、それによって、研究の時間が今まで10かかっていたのが1でできる、あるいは、研究者が実験に随分時間を使っていたのが、本当の頭のいい人は、実験室に行かなくても次のことを考えているうちにそのプロセスが進んでいきますみたいなのがAI for Scienceのありたい姿だというふうにイメージしたときに、実験施設やHPCを外に置くと、こういう基盤とか知識とかというのをどう活用するかというイメージが何か湧きづらい感じがあるんです。

今までは実験をやって、その成果をここにいずれ持ってきてやれば次のステップへ行けますよねという感じだったのが、時代が変わってきているという感に対して、全体としては旧来型でいけますよねと言っていないかと問いかけているのです。

【千葉委員】ワークフローになっている、一直線の線になっているところがそもそも古いのではないかという御指摘ですか。

【江村主査代理】そうなんです。だから、こう回っている感じの中にこの基盤が入ってくるような時代感だと思うんですね。

【千葉委員】あと、実験装置が直接AIエージェントと何かしているという絵ではないので、それも多分、気になっていらっしゃるのではないかと。

【江村主査代理】そうなんです。だから、その辺を。

【黒橋所長】NIIに絵心がなくて申し訳ないんですけども、そのグレーの線はそういうことを示しているつもりで、今日も申しあげましたように、実験ノートとか実験データはもう日々入ってきて、それをAIが処理すると。それもやるのがこのつもりなんですけれども、今おっしゃったようなループですとか、もうちょっとその拙速性がぱっと見て御理解いただけるようにアップデートはしたいと思います。

【江村主査代理】ぜひ。

【黒橋所長】ありがとうございます。

【尾上主査】吉田委員、お待たせしました。どうぞ。

【吉田委員】LLM - jp - 4になって、すごい性能が出て、すごく驚いたんですけど。

まず、テクニカルなところで興味があるのは、4段階のいろんな学習フェーズがあって、多分すごくいろんなノウハウがあって、3から4にバージョンアップしたときに何が決定的だったのかなと思うんですけど、まず、データのコーパスのサイズとか、モデルの規模とか、どの辺がクリティカルだったでしょうかという技術的な興味なんですけど、いかがでしょうか。

【黒橋所長】ありがとうございます。

8ページぐらいまで戻っていただくことはできますか。その次です。

まず総合的だという面もあるんですけども、大きなことを幾つか挙げますと、前のモデルは、この図で言えば、事前コーパスのフェーズ1のさらに3分の1ぐらいまでやっただけだったということになります。ですので、データ量は6倍になっていますし、これはもう世界的に進んできた知見ですけども、数学・コードにもっと注目するとか、QAを伸ばすという中間学習のところは合成データです。合成データで大量のQAを作ることによって指示追従性が上がると。

それから、この学習率の考え方も、従来はコサインカーブで最初から最終地点をデザインして進んでいたわけですけども、今、一定程度フラットで幾らでも試せると、その後で下げればちゃんとコサインカーブで最適化したものと変わらないとか、そういうノウハウもあります。

それから、数学の辺りのデータはかなりクオリティを上げていまして、どういうデータであればその寄与率が高いかということのアブレーションしながら進めています。

主要なものはその辺りですけども、とにかくこの図で言えば、左側のブロックのちょっとだけしかやっていなかったということが一番大きなポイントかと思えます。

【吉田委員】これ、もうコーパスも全部公開しているわけですね。

【黒橋所長】 はい。

【吉田委員】 その辺のノウハウ的なところも全部公開されているという理解でよろしいですか。

【黒橋所長】 ドキュメンテーションのスピードの問題もあるんですけども、隠すつもりは一切なくて、昨日もテクニカルレポートを書いておりますし、全て公開という考え方で進めています。

【吉田委員】 こういう本当にフルオープンというか、取組というのは、LLM - jp以外の取組って、Allen Instituteの話が出たと思うんですけど、そこまでは出るんですかね。

【黒橋所長】 そうですね。大きな規模でかなり強いモデルを出しているという意味では、もうAllen Instituteがありますので。過去にこんな取組をやってみたというところはあるんですけども、ここまで継続してずっとモデルをきちっと出しているのは、世界的にもAllen Instituteと我々だという認識です。

【吉田委員】 あと、クローズなフロンティアモデルはまだ差があるとは思いますが、もうリソースさえあればそこまで行けるみたいな手応えってあるんでしょうかという質問なんですけど、どうでしょう。

【黒橋所長】 フロンティアモデルは、恐らく計算資源が3桁違いますので、ソースさえあればという過程が成り立たない過ぎるので、あまり支えられないですかね。

【吉田委員】 なるほど。分かりました。でも、ここまで来ると、本当にAI for Scienceの大きなドライビングフォースになるし、私自身もちょっと使ってみたいなという気も。

【黒橋所長】 ありがとうございます。

今8ビリオンと30ビリオン目も多いですけども、ざっとこの10倍規模のものを今年度中に開発する。もうそれは実はアプリケーションとかは終わってしまっていて、あとは計算機さえ使えば始まるんですけども、今調達したり、使えないものが、お休みのものがあつたり、計算資源には常に困っておりますが、恐らく今年度のモデルはさらにもう一段階強くて、しっかりした使い勝手が出てくるのではないかと考えています。

【吉田委員】 サイエンスの各ドメインで独自のデータさえ作れば、そこでファインチューニングしたモデルというのは、何かすごい……。もうここまで来たら、そこでファインチューニングで、そのドメインで勝つモデルみたいなのがたくさん売れるという、そういうシナリオもあるのかなと思って。

【黒橋所長】 ありがとうございます。まさにそういうイメージでありまして、マテリアル分野のしっかりしたモデルをこれでベースでやると。

そこもちょっと議論があって、ほかのオープンウェイトモデルでもいいじゃないかとおっしゃる方もおられるんですけども、何でもともと学習しているか分からないものでいいですかというのが一つと、もう一つは、もうこういう話がしっかり進んでくれば、よその、例えば、Qwenとか出てくるかもしれない、出てきたら追加学習すればいいということですけども、このモデルはもう設計してずっと我々は開発していますので、今度やりたい追加学習というのはこんなやり方だと、そのためには学習自身も、例えば途中段階のものも含め、こんなモデルを出してくれたら、そこから我々は追加学習やりますよと。だから、半歩か数歩前からそういうコラボレーションもできるわけですね。そういうコントロール可能であるということも大きなポイントだと思っています。

【吉田委員】 はい。日本の研究者がみんなとは言わないんだけど、かなり多くの部分がクローズドなモデルではなくて、これを使うようになるような、そんな世界観ってどうやったら来るのかなと思って。ちょっとすみません、何か個人的な。

【黒橋所長】 そうですね。そういうことを目指していますし、基礎的な能力が駄目なら全然駄目なわけですけども、ここまで来れば、そういう地位につながっていく可能性は十分あるかなと思っております。

【吉田委員】 思います、私も。そうすると、収益化って、さっき後半でお金の話が出たんですけど、そこで自己収入を獲得するみたいな、そういうシナリオも幾らでも描けそうな気がしているんですね。

【黒橋所長】 ありがとうございます。

【吉田委員】 すみません。ありがとうございました。

【尾上主査】 ありがとうございます。

工藤委員、どうぞ。

【工藤委員】 ありがとうございます。大阪大学の工藤郁子と申します。

御発表いただきありがとうございました。私からはコメントが2つ3つと、質問が1つあります。まず最初のコメントは、LLMの開発について、すごく精度が上がっていて、やっぱり改めて見ると

びっくりするような勢いがある、その背後には非常にたくさんの御苦労とかがあったと思います。本当に感謝しております。

そして、今までのやり取りで示されているとおり、やはり精度が高いだけではなく、オープンであるということは非常に高い価値を持っていると思います。というのも、私は公共政策とかも一応研究をしていることになっておるんですが、フロンティアモデルが開発できる国は米中の2か国で、LLMが開発できる国は、計算資源なども加味すると、多分15か国ぐらいと言われています。その中で、日本におけるこういった取組は、その15か国以外の国々から非常に強い注目を集めているということ、公共政策の観点からは日々感じておるところでございます、18ページ目で御指摘されているように、ここで挙がっているのは韓国、インド、タイ、フランス、ブラジルという国々が挙がっておりますが、恐らくそれ以外の国々からも、ソブリンAIなどの観点から、非常に強い注目が集まるようなものだと思いますので、今後も国際的な協力を、ぜひここ以外の国々とも広げていっていただけたらうれしいなと思います。これがコメントの1つ目です。

2つ目のコメントは、同じく18ページ目にあるんですけれども、3ポツ目のところで、そのクローズドなフロンティアモデルは収益上重要なドメインや言語に集中してしまっているだけども、でも、やはり人文とか社会科学的な文化的側面というのを十分に扱う必要があるとあって、それを補う必要があるという点を強調していただいています。私は一応人文・社会科学などに属する法学とか公共政策をやっているんで、この点を指摘していただけて非常にうれしいなと思ってという点を2つ目のコメントにしたいと思います。

そして、唯一ある質問としては、利用シナリオをいくつか挙げていただいているんですけれども、この利用シナリオはどのような形で検討されたのでしょうか。翻って言うと、冒頭のほうで、後半のほうの21ページ目とかに人文・社会科学を含む学術界全体に対するオープンモデル利用環境の提供を検討しますということが書いてあるんですが、これらのシナリオは、あまり人文・社会科学にぴったり来るようなものではないような気がしました。もちろん、人文・社会科学と一口に言っても、実験をやるものもあれば実験をやらないものもあるので何とも言えないんですけど、何かぴったり来るような感じもしないなと思ったので、どういった検討過程でこのシナリオを開発したのかということのをちょっとお伺いできればと思いました。いかがでしょうか。

【黒橋所長】まず、いろいろ励ましのコメントをありがとうございます。

御質問ですけれども、これはNIIのセンターでディスカッションして、確かにちょっと我々、自然科学研究者的な視点が強いかもしれないかもしれませんけれども、でも、ROISといいますか、NIIとしては、人間文化機構とか、あるいは、デジタルヒューマニティとか、そういう関係もかなりございますので、そういうところは一定程度は意識したつもりであります。例えば、古典籍の分析ですとか、そういうことも含めてですね。

至らない点もあると思いますので、それはぜひディスカッションというか、ここにシナリオを加えさせていただければと思っております。

【工藤委員】ありがとうございます。

それに対する追加のコメントで、途中で言及もございましたAI for Scienceに関する助成プログラム、チャレンジ型だったものが.....。

【黒橋所長】SPReAD。

【工藤委員】SPReADって、多分そんなに多くないかもしれないんですけど、一部人文・社会科学からももしかしたら応募があるかもしれないので、そういったものも反映していただきつつ、おっしゃっていただいたとおり、現場の研究者との対話とかフィードバックとかも今後やっていただけると非常にうれしいなと思いました。ありがとうございます。

最後のコメントは、こちらの黒橋先生の御発表とか、NIIさんのお取組の話ではなくて、今後の取りまとめに向けての思いついたことをしゃべるというコメントが一つございます。

お話を伺っていて非常に心強く感じたというか、今後AIやLLMによって研究が支援されて効率化していくというイメージがだんだん湧いてきたんですが、効率化とイノベーションはイコールではない気がしています。AI for Scienceの推進に向けた基本的な戦略方針にいつも書いてある三角形の図の一番上の「世界を先導する科学研究成果の創出」というところは、やっぱりイコールではないような気がしています。すごく卑近な素朴な実感からすると、今までやってきたような研究の方法とか研究の進め方というのが、AIやLLMやデータの支援によって非常に短くなっても、空いた時間に、例えば、先ほどの例で言うと「こういう分野の研究者と会いに行くと学際的でいい共同研究ができるかもしれない」ということをAIに教えてもらった後に、ちゃんと交流する時間が確保できるかとかが多分本当はイノベーションとかに関わってくるところだと思うんです。けれど、AIが効率化して空いた時間

を私たちは本当に研究に注ぎ込めるかということ、かなり怪しいというか、教育もあるし、学務もあるし、公務もあるし、その他いろいろな事務作業とかもあります。なので、全体として調整していかないと、研究の中核だと思われるコアな部分だけ効率化しても、本当にこの三角形の図は上に行けるのかみたいなのがちょっとイメージがつかないので、また別の考慮が必要なのではないかと思えます。あるいは、ほかのワーキンググループとかほかの研究会で考えられている戦略とかとセットで、両輪として考えていかないと、ちょっと難しい。せっかく効率化しても科学者の手が空かないんじゃないかみたいな気がしており、この点をまた取りまとめに向けてコメントとして残したいと思いました。

以上です。ありがとうございます。

【尾上主査】ありがとうございます。

矢守委員、どうぞ。

【矢守委員】私からも、工藤先生と少し重複するところがあるんですが、まずは、国産のNIIさんの作成されたLLM、AIに関しましては、非常にすばらしいなと思って拝聴しておりました。非常に中身の濃いといえますか、性能の高いものを作り上げられたのかなと思います。関係者、お疲れさまでした。

私からは、大きくコメントになります。やはりこのように、どういうふうにしたか、中のデータも併せて分かるようなAIを我々が持つということは、学術的な意味を思っても、非常に重要なことかと思えます。

やはり研究者としては、様々なAIを研究にもう活用しているという段階に至っておりますが、今後これがこのままの形で使えるようになるかということに一抔の不安を覚えております。

そのような中で、やはり我々がコントロールできるAIがあるということは、今後の学術研究においても非常に大きな意味があると個人的には思っています。その意味で、先ほど学際的なというお話もありましたけれども、今度の取組でやられているところで様々なデータを集めていくというのも非常に重要かと思えますが、できることからやっていくという考え方もいいのかなと思っています。やはり自然科学系のデータというもののほうが現状集めやすいですし、使いやすいものになっているのは間違いありません。そこのところで作り上げていきながら、人文科学系、それから、なかなかデータ化しにくいものも、むしろどうやってデータ化していくのかということも併せて、多分研究の裾野が広がっていくのではないかと思います。

そういう意味で、この波及効果といえますか、シナジー効果というんですか、今までAIとちょっと遠いなと思っていた研究者とも、本当の意味での学際研究ができる可能性を持っている、そういうポテンシャルを感じるAIかなと思いました。

一つ質問なんですけれども、やはりこれは宣伝が大事かなと思っています。使える方から使っていくということからスタートにして、そのステップを踏みながら、やはり効果があるということと、使いやすいところを、利用者もしくは潜在的に利用者になる方々にどうやって使っていけばいいかということから広めていくのが一つのステップかなと思うんですが、何か現状でそういったことをお考えになっていたりするのでしょうか。もし既にお話しになっているようでしたら申し訳ありません。ちょっと聞き逃してしまったかもしれませんので、お願いします。

【黒橋所長】ありがとうございます。

これも温かいといえますか、オープンな意義を御認識いただいて本当にありがとうございます。これをいろんなところで説明するのに我々も本当に苦労しておりますので、今のような御意見は大変ありがたく思っております。

それから、途中の部分になりますけれども、オープンであることが、今クローズであるために著作権者との信頼性のなさみたいなのがあって、それで逆に良いデータがなかなか使えないという面もあります。そういう意味で、オープンにすることによって、逆に自分のデータは学習してもいいんだよということの世界も開けていくと思いますし、人文・社会科学のデータも、理解いただき入れていただければと思っています。

宣伝につきましては、これもまさに数日前に東京大学の方と相談をさせていただいたところですが、GakuNinの認証があれば、先ほどのOpenWebUIのインタフェース、このモデルを使っていただけということのテストサービスは近々に始めようと思っております。今こちらにおられる東大・千葉センター長もうなずいておられますので、まだモデルはここまでは来ましたが、恐らくChatGPTをしっかりと使っている方からすると、もうちょっとということもあるかもしれませんし、あるいは、大規模な運用をしていくということもいろんな課題はあるかもしれませんが、おっしゃっていただきましたとおり、まずとにかくスタートするんだという考え方で進めたいと

思っております。

以上です。

【矢守委員】ありがとうございました。

【尾上主査】石田委員、どうぞ。

【石田委員】九州大学の石田です。御説明いただきましてありがとうございました。

御説明いただいたようなAI利用環境ができれば、本当に詳しくない人でも使えるようになるので、非常に魅力的だなというふうに思いました。

私、研究データの流通とか利活用の立場ですので、その点からお聞きしたいことなんですけれども、これ、まさに先ほどの絵の問題なのかもしれませんが、例えば、この知識基盤で今御説明なさった環境と、そこで作り出されるデータ、そういったものをまた吸い上げてモデルの学習などに使うというような流れみたいなものは御検討なさっているのかということが一つ質問と、もう一つ、その場合、もちろん全てが学習に利用できるデータにはならないかもしれませんが、学習できるデータの中でも、データの質とかタイプとか、いろんなものがあると思うんですけれども、そういうものでもモデルの学習には有効なのかどうかというところがちょっと私、分からないもので、教えていただければと思います。

以上です。

【黒橋所長】ありがとうございます。

まず、そういうサイクルをまさにつくっていくのが、NIIとしていいですか、国として重要なことだと思っております。

このよくない絵のNII RDCのところには、実は管理・共有する基盤で、まだ発表前の研究データなどを共同研究者できちっと管理しながら、そこで研究を進めていくというフェーズのものと、それから、JAIRO Cloudという、一旦区切りがついたもので、ここまでは公開できるというものを公開するような基盤もございます。そこでは、その研究グループだけではなくて、世界中の人がその研究データを活用することもできるわけですけれども、そこで出てきたデータのあるドメインの研究者が、このデータも使って自分のモデルをもっとよくしていこうという、そういう追加学習ができる環境だと。それは研究グループぐらいの単位かもしれませんが、もしかしたらマテリアル全体でこれとこれを足した、いいモデルをつくらうよということになれば、それで作られるかもしれないので、そういうサイクルができてくるということは、まさに目指していることです。

それから、御指摘のデータのクオリティの問題も、それも本当にありがとうございますなんですが、そこをきちっと来歴が管理できるのがこの基盤の強みでありまして、クロールしてきて何かデータを探してくることはできるんですけれども、それとはちょっとレベルの違うきちとしたデータのクオリティ管理があるので、ここのグループが作ったものは使おうとか、論文でこういう良さが言われているならば使おうとか、そういうことをきちっと把握してモデル学習なんかフィードバックできる、そういう情報のついた基盤になっているといいですか、それをどんどん進めていきたいということです。

一般的には、特にドメイン特化の場合には、逆に、ここまでのLLMは少々ノイズがあっても本当に大規模で、だんだん能力、理解力は上がっていきますけれども、最後ドメイン特化する場合には、本当にクオリティが重要だと思います。ですので、その意味で、来歴がきちっと管理されているという基盤があることの意味は非常に大きいと思っております。

【石田委員】ありがとうございます。

【尾上主査】千葉委員、どうぞ。

【千葉委員】東大の千葉です。

いろんな御意見が出たので、もうあまり言うことはないんですけれども、ここまでモデルがいいものができると、やはりその先のAIエージェントをどうやってつくっていくかというところにもう検討できるステージまで我々は来たのではないかと考えていまして、先ほど既にDeep Researchを組み込んで動かす実験はもう始めているとおっしゃっていたんですが、これからはどんどんいろいろな分野ごとに、先ほどフィジカルAIの話も出ましたけれども、それもある種のエージェントですし、実世界とつなげていくというところをどんどん作り込んでいかなければいけない時代に来ていると思うので、もしそういう取組が既に行われているのでしたら御紹介いただきたいのと、こういった場ですので、私たちはそういうことも考えなければいけませんねということコメントさせていただきたいと思っております。

以上です。

【黒橋所長】ありがとうございます。

まず、エージェント化していくということもそうですし、とにかくやっここまで来た。ある種の地頭の賢さが重要で、そこまで来れば、実は今日ちゃんと申し上げていない気がするんですけども、強化学習はまだしていません。なぜなら、強化学習するためのベースの賢さが必要で、前のモデルはそのレベルに行っていなかった。今回は選択学習みたいなことはしていますけれども、やっ強化学習してよくなりそうなところまで来たので、それをきちっとやって、さらに改善すると思っています。

同じことが、先ほどちょっと御紹介したDeep Researchがもうできるようになったとか、そういうことにもなっていますし、AIエージェント化していくため、ツールコーディングなどができるためのファインチューニングのデータの構築もずっと進めてきておりましたし、それも今回強化していこうと思っています。そういうことで、エージェントとしてしっかり動くという能力もどんどん上げていきたいと。

最後、御指摘のあったロボティクスと実世界等ですけれども、LLM - jpの活動の中には、実環境グループというグループもありまして、当初からというか、早い段階から、尾形先生にもその主査を務めていただいて、そういう活動も進めておりますし、それから、皆様も御存じだと思いますけれど、経産省さんが進めようとしているフィジカルAIのプロジェクトともいろいろコラボレーションを検討しておりますので、この今のモデルがしっかりしたものができつつあるということを中心に、今おっしゃったような拡張モジュールを進めていきたいと思っています。

ありがとうございます。

【尾上主査】江村委員、どうぞ。

【江村主査代理】ちょっと質疑の中で出てきたことの関係で、一つが、産業界の利用みたいな話がちょこっとあって、先ほどGakuNinでスタートしていけばいろんなことができますとおっしゃっていて、そこをどこまで展開するかというのが、実は広げていく上では結構重要なポイントかなと思っているんですけど、そこはどうお考えになっているのかというのが一点と、もう一つが、多分メタデータの話と関係するのかもしれませんが、この対話型サーバという中で、アイデア出しとか分野横断というのがあって、機能としてそれがインプリされていると思うんですけど、結局分野融合がこの国は起きていないという問題に対して、仕組みだけではなくて、融合を喚起するような工夫をしていかないと難しい部分があるかなと思っています、その辺について何かお考えがあればお伺いしたいと。

【黒橋所長】ありがとうございます。

まず1点目の認証については、これは常にNIIがいろいろなところで発表しても御指摘いただいている点、それから、産業界との連携の重要さもありまして、スタートとしてGakuNinに認証してアカデミアで使っていただくということは、まず本当の立上げとして始めたいと思っていますけれども、例えば、まずはアカデミアと協力されている企業の方ですとか、さらにもう少し広げた範囲にこの認証機能をどうやって拡張していくかということは常に議論しておりますし、いろいろ具体的な計画も持っておりますので、もしあれなら後で補足いただければと思います。

2点目の分野融合を本当に進めるにはどうするかという話ですけれども、一つ、やっぱりAIの強みは、これもう今日本での医師国家試験に通るレベルの能力を持っていますし、物理学も化学も全部分かっているような、まずファシリテートできるレベルの賢さをもう既に持ってしまったと。先ほど共同研究者を提案できますよみたいなことを申し上げましたけれども、もっと積極的に、あなたの研究はこういう研究で、まずそのドメインとしてすばらしいです、だけどこういう社会課題があるので、こういうことに挑戦してみたらどうですか、そのときにはこういう組合せが考えられますよと。皆さんが例えばCRESTを考えると、じゃ誰と協働しようかみたいな考え。それは、でも、少し範囲がそこまで広くないと思うんですけど、もっともっと広い範囲の人が一緒になって、例えば、水の問題とか、地方の問題とか、地域紛争の問題に取り組むための提案をしていく、そういう機能をAIに持たせることもはやできるようになってきているというか、そこが重要かなと思っています。しかし、時間がないみたいな話は、また別の問題ですけれども。

【江村主査代理】だから、やっぱり使う側の意識が、それを活用できる側に行かないと。

【黒橋所長】最後はそこだと思うんですけど、ただ、それを探すことすら今までは極めて難しかったのが、一段階は上がるんだと思います。

もし認証について補足いただければ。

【合田副所長】技術的な点の補足をさせていただきますと、おっしゃるとおり産学連携は重要で、やっぱりSINETにつきましても、産学連携している企業はもう利用できたりですとか、NII RDCについても、もちろん産学連携プロジェクトの利用が始まっています。

認証というのは、やはりデータのアクセスにおいて非常に肝となる部分でありまして、これは単に

ID パスワードが一致したからいいという世界ではなくて、その人が本当にその企業に所属している、学に所属していることを担保した上で利用する必要がありますので、慎重に進める必要があると。

技術的には2つの解をつけて進めています。1つは、産業界の持っているIDとGakuNinのIDをひもづけるようなプロキシする方法を提供するもの、もう一つは、産業界もいろいろな使い方がありますので、そもそも産業界としてIDがなかなか我々の求めるレベルのものがない場合もありますので、それについては、学の側がそのIDも面倒見ることも含めて、今技術的な検討ですとか実証を進めているところであります。

【江村主査代理】ありがとうございます。

【尾上主査】ありがとうございます。

1点、私から、無茶を承知でお伺いしたいところなんですけれども、これは前回の山地先生の御発表のときにも申し上げたんですけれども、第7期の基本計画の期間にもう入っている我々からすると、この将来像の基盤って今すぐに欲しいような状況だと思っています。例えば、先ほども黒橋所長がおっしゃっていたように、SPReAD、ARiSE等が進んでいって、これで、先ほどの江村委員の話等もあるのかもしれませんが、科学研究の核心と言っているぐらいなので、研究スタイルが大きく変わって、要するに、学術研究プラットフォームの使い方が変わってくるということは、これは織り込みながら多分我々としてはこのワーキンググループの取りまとめもしていく必要がございますし、今日も黒橋先生に出していただきましたけれども、そのための数値的な裏づけというのも一部入れていく必要があるかなと思っています。

一方で、先ほどもお話しいただいたように、知識基盤のベースとなるものが出来上がっていない状況でそれを議論してもしょうがなく、ようやくそこまで行って、じゃ、これから例えば今の科学研究革新プログラムの成果が出てきたときに、そこにどういふふうにこれをアジャストして対応させていくかというようなところまで、多分在り方としては、我々としては、そこもある程度フォアキャストみたいな形で入れながらやっていくといいのかなというふうに思っているんですが。

全く読めない未来予想をお聞きしたいんですけれども、例えば、先ほどのSPReAD、ARiSE等がこれから1年、3年やってきた結果で、先ほど後ろの後半で出していただいたような数値というところが、研究スタイルが変わっていくと、どれぐらい上振れ、下振れはないと思うんですけれども、上振れする可能性があるのかというようなところが何か分かると、そういう情報も教えていただけると、我々としては、こういう中に入れ込んでいけるかなと思ったんですが。難しい無茶振りと言って申し訳ないんですが、いかがでしょうか。

【黒橋所長】それは本当に難しい御質問だと思います。全てのものが加速度的に変化しておりますので、今まさにおっしゃったとおり、上振れがどういう加速度で起こるかというのは本当に難しいところなんですけれども、やはり人文・社会科学の方も含めて使われるようになるということが期待されますが、また、そのときの処理量が大规模データ処理よりはもう少しは少ないのかもしれませんが、その辺りも含めて、本当にそもそも分野の多様性があるって、今の利用率があるって、どこがぐっと伸びて、しかし、そこは本当にさらに大きな計算資源を必要とする分野なのかということも含めると、ちょっとすみません、簡単には申し上げられませんが、SPReAD、ARiSEが立ち上がって半年ぐらいしたら、そんな悠長なことは言っていられないのは本当に私もそう思いますし、スピード感が重要ですけれども、もう早い段階からこういう試算をいかにアップデートしていくかということをきちっと考えることは重要なことだと思います。

すみません。それは数値的にはなかなか難しくて。

【尾上主査】ありがとうございます。

先ほどおっしゃっていただいたように、ここを出していただいた数値を多分明らかに上振れで上に超えていくような数値が必要だということと、あとは、先ほどの今走っているプログラムからいかにそれを早くフィードバックして、ここの数値をアップデートしていくということが多分国の施策としては必要になってくるという。

【黒橋所長】すみません。もう一個、最初に申し上げて。

やっぱりそういう調査も必要だと思っていて、そういう場合は、例えば学術会議だと思っただけですね。学術会議に対して、これからAI for Scienceでどれぐらいそれぞれの分野が変わっていくか、スピード感があるかということ調査して、それを一つの根拠としていくということは重要だと思いますが、学術会議の役割だと思っておりますし、学術会議はもっと社会にとってまさにその重要な役割の一つかなと思っております。

【尾上主査】ありがとうございました。

千葉委員、どうぞ。

【千葉委員】補足なんですけれども、私、さっきAIエージェントの話を持ち出させていただいたんですが、その理由は、AIエージェントが動き出すと、いわゆるモデルに対する問合せもAIが出すようになるので、残念ながら、残念じゃないかもしれないんですけれども、スピードは桁で上がっていく可能性があって。今、例えば、私、研究分野はソフトウェアなんですけど、今学生さんたちは、もうエージェントを自分の部下のように使って動かすので、それらはすごい勢いでモデルに問合せをして、書いて直して、書いて直してという感じなので、やっぱり一桁二桁上がっていく未来が見えるのではないのでしょうか。

それは今はソフトウェアの分野だけですけれども、ロボティクスに本格的に応用されると、恐らくはあちらもそうなるでしょうし、ほかの分野、それこそ人文・社会の分野も、大ざっぱな指示を与えると、まずAIが計画を立てて、子分のAIにさらにやれという、AIエージェントに投げて、AIエージェントがすごいスピードでモデルに問合せをやって、分岐は何だ、次の分岐は何だとなるので、ちょっと想像がつかないというのは、結論は黒橋先生と同じなんです。

すみません。以上です。

【尾上主査】ありがとうございます。無茶な質問でございました。

【千葉委員】そこを目指さないといけないということだと思っています。

【尾上主査】いや、おっしゃるとおりです。

そろそろお時間になってまいりました。まだいろいろ御意見あるかと思いますが、またそれらについては別途いろいろ集める方法を考えたいと思います。

それでは、本日の議論で出た皆様のコメントなども踏まえまして、次回のワーキンググループでは、取りまとめの素案の議論に持っていければなと思っています。

続いて、議事(2)のその他ですが、事務局より、御紹介事項などあればお願いいたします。

【麻沼参事官補佐】事務局でございます。

資料3を御覧ください。ワーキンググループの今後の進め方についての案という資料でございます。資料1の中でも今後のスケジュールを少し御紹介しておりましたが、6月中には取りまとめを行いたいということを申し上げました。次回は第5回になりますが、5月22日金曜日を予定しております。その次が6月16日火曜日を予定しております。

次回、第5回ですけれども、これまでは国立情報学研究所からの構想について御議論をいただいていたところですが、ユーザ側、連携する側からの御意見も取り入れたほうがよいのではないかと考えておりますので、主要分野における研究データの管理・利活用状況などについてまた御議論をいただければなと考えているところでございます。また、取りまとめの素案についても御提示をさせていただきまして、議論に入っていければと考えております。

第6回目、6月16日のほうは、取りまとめ案について審議をしていただく予定でございますけれども、終了しない可能性もございますので、その場合は、必要に応じてメール審議などもさせていただければと考えておりますので、御協力をいただければと思います。

各委員の先生方におかれましては、今の御予定のほう、5月22日金曜日16時から18時、6月16日火曜日の16時から18時を、御予定は仮押さえをさせていただいているかと思いますが、引き続き押さえさせていただくようお願いいたします。

事務局からは以上でございます。

【尾上主査】ありがとうございます。

委員の皆様、取りまとめに向けて、引き続き御意見等を賜ればと思いますので、どうぞよろしくお願ひ申し上げます。

最後に、事務局からありますか。お願いいたします。

【麻沼参事官補佐】本日も御議論いただきましてありがとうございます。

取りまとめに向けまして、本日も少し御意見いただけたところですが、メールでも御意見頂戴できればと思っていますので、後ほど事務局よりメールをお送りさせていただきますので、本日も言い足りなかったこととか、取りまとめに向けた骨子案への御意見でも結構でございますので、いただければと思います。どうぞよろしくお願ひいたします。

以上です。

【尾上主査】ありがとうございます。

それでは、本日の議題はこれまでとなりますので、これにて閉会とさせていただきます。どうもありがとうございます。次回もよろしくお願ひ申し上げます。

— 了 —

お問合せ先

研究振興局参事官（情報担当）付学術基盤整備室

（研究振興局参事官（情報担当）付学術基盤整備室）

ページの先頭に戻る

文部科学省ホームページトップへ