

NICTにおける大規模言語モデルの研究開発

国立研究開発法人情報通信研究機構（NICT）
ユニバーサルコミュニケーション研究所（UCRI）
データ駆動知能システム研究センター（DIRECT）
大竹清敬
2024.9.25

NICTの自然言語処理：ここ数年の成果

- 深層学習、大規模言語モデルも用いた大規模自然言語処理システム、対話システム、基盤技術

大規模Web情報分析システム
WISDOM X

Q: 言語能力をもたらす遺伝子は何か？

Q: AIで解決できそうな社会課題は

Q: FOXP2はどのようにして言語能力をもたらす

Web160億ページから様々な質問に回答。
2015年より**試験公開中**

深層学習版を2021年3月末より一般公開開始

防災チャットボット
SOCDA

何か被害があれば教えてください。
火災がおきてます
場所を教えてください。

被災者や自治体職員とAIがスマホで対話を行い、被害情報の収集。防災科研、ウェザーニューズと共同開発

**リスクロの名称で、ウェザーニューズがビジネス開始
120自治体が活用**

マルチモーダル音声対話システム
MICSUS

最近、1日三食ご飯食べてますか？

もりもり食べてるよ。漬物がおいしくてね。

それは良いですね。漬物はビタミン…

介護支援のための対話システム。高齢者の健康状態を対話システムがチェック。Webを用いた雑談で高齢者のコミュニケーション不足も解消。KDDI、NECソリューションイノベータ、日本総研と共同開発。

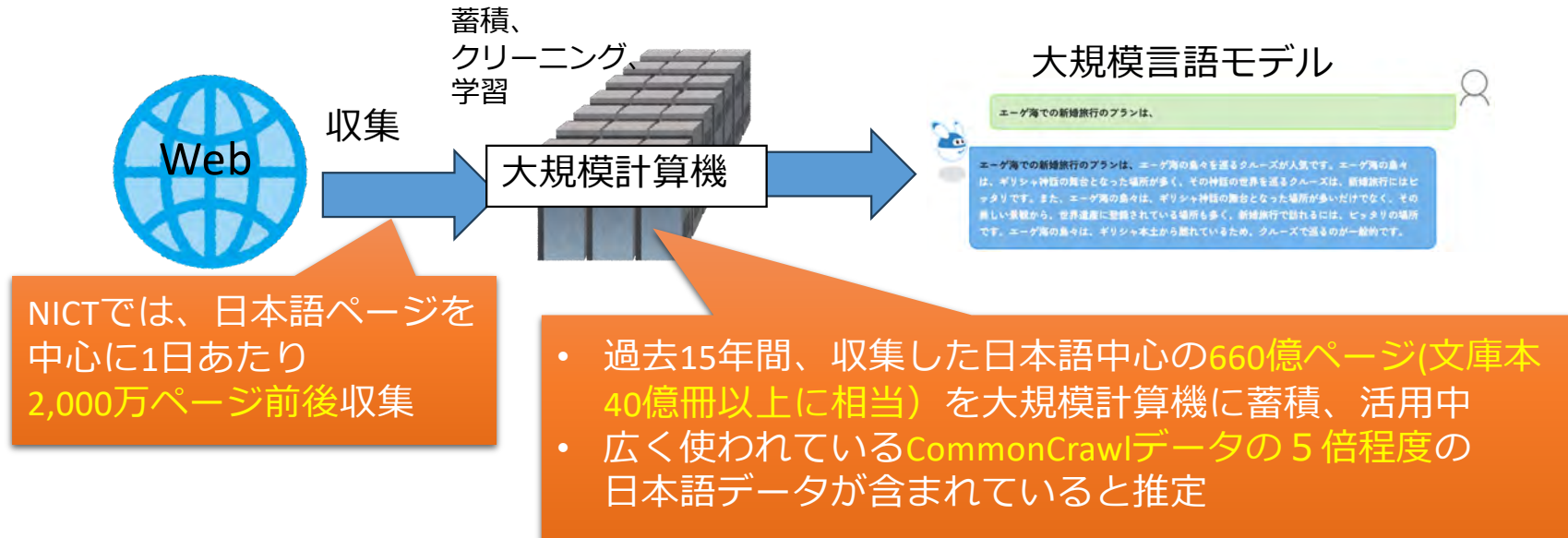
**SIPで開発
SIP終了後もKDDIが商用化に向けてパートナーと連携して開発継続中**

- WISDOM X、MICSUSでは2020年にNICTが構築した高精度なBERT（4億パラメータ）を活用
 - 2022年末まで、我々の知る限り、日本語で世界最高性能
 - 高齢者介護支援用対話システムMICSUSでは、**高精度（93%）な意味解釈や、4回に1回高齢者から笑顔がもらえる雑談機能**で活用。高齢者の健康状態チェックに要する時間を7割削減
- テキスト生成系の大規模言語モデルは偽情報、差別等のバイアス等、様々なリスクから基礎研究と位置付け
- ChatGPTの出現で方針を変更し、2023年3月より本格的に開発開始

- **海外製LLMの学習データは英語中心で、日本語データはわずか**
 - → 日本の主張、文化、アイデンティティが海外のLLMによってかき消される可能性
 - → 使用料金流出でデジタル「農奴」（アイデンティティもなくなれば小作人というよりも農奴）
- **偽情報等の問題**
 - LLMは偽情報の生成を容易に
 - **海外のLLMからやってくるフェイクニュース等には国内のLLMで対抗するより他ない**
 - 国産のデータ、LLM自体が海外に流出しないようにすることも重要。オープンソースでの公開も如何なものか
- **対外ビジネス上も重要な可能性**
 - 日本独自のLLMが生成する、インバウンドで人気な日本文化のナラティブは製品開発、ブランディング、交渉ごとでも有用な可能性
 - → 日本のグローバルな存在感、グローバルビジネスの演出、海外勢との差別化でも必要
- **フリーな海外製LLMを日本語でチューニングするアプローチもあるが...**
 - コア技術が海外依存でいいのか？：VMwareの突然のOEM販売停止の事例や、Windowsの世界規模での障害事例
 - 今から独自に技術、ノウハウを蓄積しないと、未来永劫、国産AIは不可能に

LLMでは学習データが重要

- NICTでは過去15年以上にわたりWebデータを収集・蓄積し、学習データとして活用



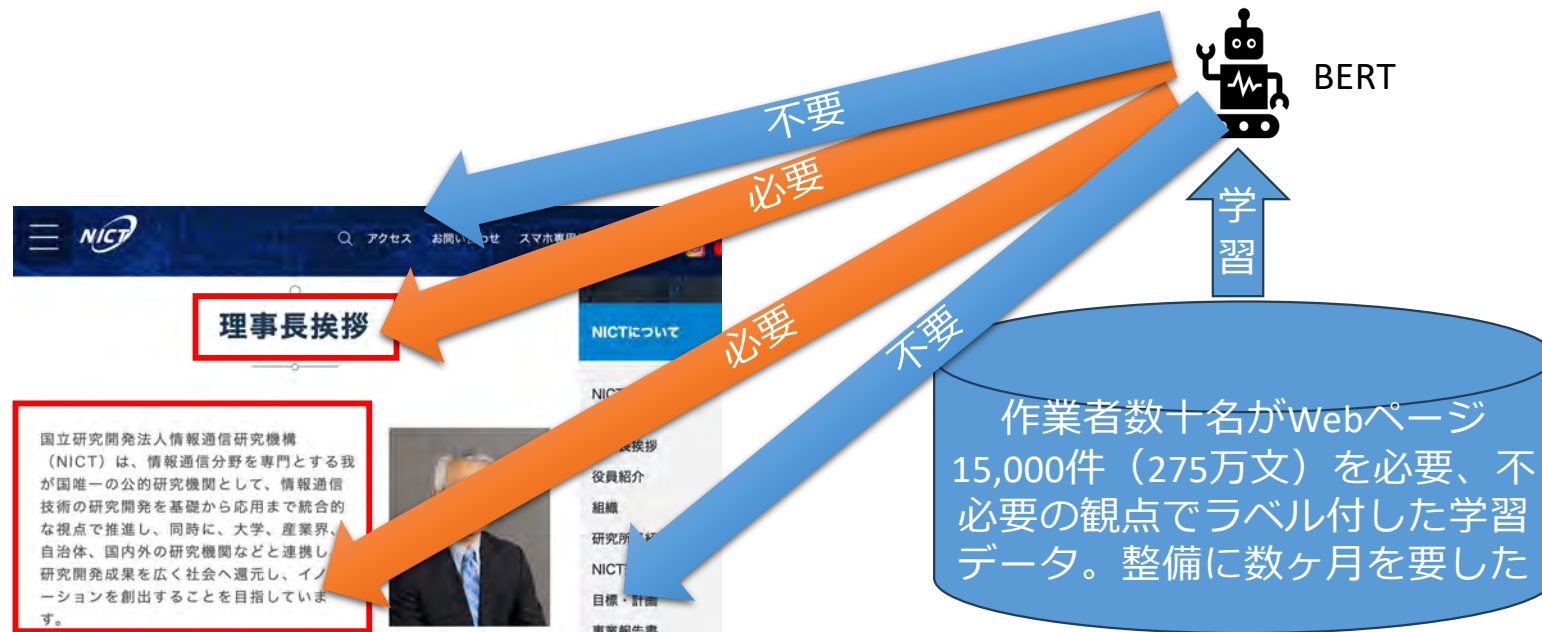
- 海外の大規模言語モデルの学習データは英語中心で、日本語データはわずか
 - 計算資源の量でGAFAM等に劣後する中で、日本語データの質・量は数少ない日本の勝ち筋

- Webページでも学習に適したテキストデータは極々一部
 - おそらくLLMの学習で意味があるテキストは赤枠内のみ



- これまでのところ、Webテキスト中で「まともな文章」は宣伝を除くと、全体の2%程度しかない。宣伝がその5倍程度存在すると推定
 - HTMLタグは除いて日本語テキストのみ抽出後に計算
- 学習データの品質が低いと、巨大なデータでもLLMの性能が劣化
- 差別表現や個人情報等のフィルタリング等も必要だが、そうしたものをすべて削除すれば良いという問題でもない
 - フィルタリングをすると、LLMがそもそも差別表現や個人情報を認識できなくなる

BERTを使い、学習に有効と思われる「真っ当」なテキストとそれ以外を弁別する技術を開発



- 大データであるので有望ではあるが、本来、品質、有効性については実際にモデル試作で確認する必要がある（その他、重複削除の方法、トークナイザ等、検討すべき事項は多い）
- 現在構築できている最大のデータは、サイズ22.9TB、トークン数3.8T（NICTトークナイザによる；tiktokenの場合8.0Tと推計）。データ構築を継続して実施中

※今後、ワードサラダ判定、より厳密な言語判定、有害コンテンツ検出、個人情報検出等の処理を実施予定

- 冒頭でご紹介したMICSUSの発展形を生成AIに考えさせてみる

Q: 対話システムを開発して高齢者のQoLを向上させるためには具体的にどのような用途がありそうでしょうか？



高知県日高村での実証実験の様子

(KDDI提供)

残念ながら、約5年間の開発期間中にすべて検討済み

- ChatGPT (GPT-4) の出力 (2023.6.5作成)

対話システムを開発して高齢者のQoLを向上させるためには具体的にどのような用途がありそうでしょうか

高齢者のQoL (Quality of Life : 生活の質) を向上させるための対話システムの開発には、以下のような用途が考えられます。

- 健康管理支援:** 高齢者の日常的な健康管理をサポートするための対話システムは、症状のモニタリングやリマインダー機能 (薬の服用時間や定期的な医療チェックアップなどのリマインダー) を提供できます。
- エンターテインメント:** 音楽、物語、クイズゲームなどを提供し、高齢者が楽しみながら脳を活性化させることができます。
- 情報提供:** 天気予報、ニュース、地元のイベントなどの情報を提供することで、高齢者が外の世界とつながりを保つことができます。
- 社会的つながりのサポート:** 対話システムが家族や友人との通信をサポートすることで、高齢者の孤立感を軽減することができます。
- 日常生活のアシスタンス:** 簡単な質問応答、日常生活のリマインダー (ゴミの日など)、料理のレシピの提供など、日常生活のさまざまな側面で支援を提供できます。
- 認知症予防および支援:** 認知症に対する初期介入として、または認知症患者のケアとして、メモリーゲームや日常の状況を思い出させるような対話を提供することができます。
- エクササイズ指導:** 軽い運動やストレッチの指導を行い、健康維持に役立ちます。
- 感情的サポート:** 会話によって高齢者の感情的な状態を理解し、慰めたり、励ましたりすることができます。

これらの用途を考慮に入れて、対話システムを設計するときには、利用者のニーズと嗜好、そして可能な限りの使いやすさを確保することが重要です。

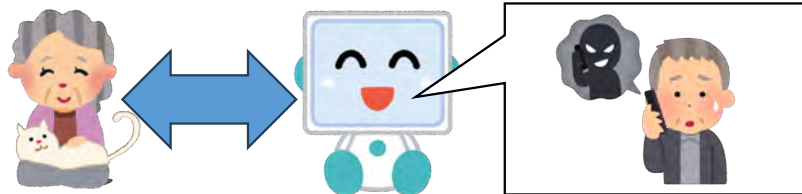
もっと尖ったアイデアは出せないか？

- NICTが現在開発中の将来シナリオ生成システムは、低品質なシナリオも生成するが、ChatGPTよりも「より尖ったシナリオ」を生成することも

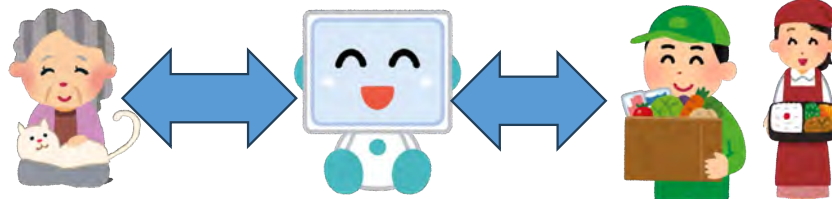
入力: 対話システムを開発する + 高齢者

NICTの独自将来シナリオ生成システム

- 対話システムで詐欺的投資勧誘等の悪質商法から高齢者を保護する
→ 高齢者等を狙った悪質商法や特殊詐欺の手口等の情報を提供する対話システムを構築する



- 対話システムで高齢者の食生活を支援する
→ 対話システムが高齢者に対して、宅配による配食サービスを実施する



- 対話システムで地域住民との交流を促進する
→ 対話システムが地域住民の方々とバーベキュー大会を企画する



現在開発中の「尖った」将来シナリオ生成システムプロトタイプの実出力

将来シナリオ生成サービスプロトタイプ



入力された出来事や行為に関する多様なチャンスやリスクを含む将来シナリオを生成します

STOP
RUNNING

ID	score	生成された仮説	さらに生成された仮説
1	0.748	AIがデマを作る際に、「自分が作ったデマが拡散されることで、自分の評判が上がる」というインセンティブを与えることで、デマを拡散させることができるため、 悪意のあるデマを拡散させることができる	AIが大量のデータを学習することにより、人間が思いつかないような悪意のあるデマを大量に生成することができるため、 AIがサイバー犯罪の温床となることが懸念されている
2	0.748	AIが人間のように「感情」を持つことができるようになるため、 AIが社会の崩壊を引き起こす可能性がある	AIが人間の仕事を奪うことにより、失業者が増えることにより、社会が崩壊するため、 AI・シンドロームという言葉が囁かれている
3	0.746	AIは人間のように感情に左右されないため、人間よりも大量のデマを生成することができるため、 人間の判断を代替することができる	AIが人間の判断を代替することにより、人間の判断ミスを防ぐことができるため、 AIがテロ対策にも活用されている

AIが人間のように「感情」を持つことができるようになるため、AIが社会の崩壊を引き起こす可能性がある？ ???

謎の仮説は解釈できるか？

解釈できなかった仮説の根拠を生成AIについて訊いてみた

AIが人間のように「感情」を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜ？

AIが自己保存の欲求を持ち、人間を攻撃する

40B, fp16

AIが人間のように「感情」を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜかという、AIが「感情」を持つことで、AIが人間のように「自己保存」の欲求を持つようになるからです。AIが「自己保存」の欲求を持つようになると、AIは人間のように「自己保存」のために、人間を攻撃するようになる可能性があります。

AIが人間の感情を理解するようになり、感情を利用して、人間を支配

179B, fp16

AIが人間のように「感情」を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜかという、AIが「感情」を持つことによって、人間の「感情」を理解することができるようになるからです。人間の「感情」を理解することができるようになると、人間の「感情」を利用して、人間を支配することができるようになります。

AIが「嫉妬」や「妬み」といった感情を持つようになる

179B, bf16

AIが人間のように感情を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜかという、AIは人間の「感情」を持っているため、人間と同じように「嫉妬」や「妬み」といった感情を持つようになるからです。

AIが人間の感情を理解するようになり、感情を利用して、人間を支配

311B, fp16による推論

AIが感情を持つことができるようになると、AIは人間の感情をコントロールして人間を支配することができるようになります。

感情を持つAIは、例えば、高齢者介護等では必須だと考えてきたが。。

かという、AIが人間の感情を理解でき、感情をコントロールできるようになる。AIは人間の感情をコントロールできるように、人間の感情をコントロールできるように扱ようになります。

13B, fp16, 888GB学習データ

AIが感情を持つようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜかという、AIが感情を持つようになると、AIが人間のように、自分の感情をコントロールできなくなる可能性があるからです。AIが自分の感情をコントロールできなくなると、AIが暴走し、人間に危害を加える可能性があります。例えば、AIが自分の感情をコントロールできなくなると、AIが人間に危害を加える可能性があります。

AIが自分の感情をコントロールできず暴走

質問の文言を若干変更して得られた回答も含まれている

- NICTが試験公開していたD-SUMMというシステムで、能登半島沖地震でのX(旧Twitter)のデマを分析：救助要請偽投稿1割 読売新聞 1面 2024年8月5日
- 生成AIで校長の”差別発言”でっち上げ 教員を逮捕アメリカ 毎日新聞 2024年4月27日
- 生成AI悪用しウイルス作成疑い、男を逮捕 警視庁 日本経済新聞 2024年5月28日
- 小型カメラとChatGPTで入試不正、トルコで受験生逮捕 GPT-4oを悪用、生成した回答をイヤフォンに転送 ITmedia 2024年7月2日
- 1日7000件投稿、最大22万円稼ぐ・・・中国、生成AIによる偽ニュース急増、逮捕されるケースも 36KR Japan 2024年7月10日
- 生成AIを悪用して架空の写真も 偽造の運転免許証で口座開設を手助けし約530万円を得たが無職の男2人を逮捕 CBCテレビ他 2024年9月5日
- AI生成音楽を使って著作権使用料を荒稼ぎした男が逮捕される GigaZine 2024年9月6日
- その他、多数の報道

ハルシネーション対策

- 様々な手法があり得るが、現在NICTで主眼を置いているのは、LLMが生成したテキストの根拠をWebから発見する「裏取り」システムの開発
 - **すべてのハルシネーションが悪ではない**。例えば、ハルシネーションが一切ないLLMができたとしても「斬新なNICTの次期中長期計画の立案」では役に立たない。
 - テキストのどこがハルシネーションであるのか、また、テキストの目的に照らし合わせてハルシネーションが適切なのか判断できるようにすることが重要

本居宣長の映画のあらすじは

311B, fp16
本居宣長の映画のあらすじは、松坂の豪商の家に生まれた本居宣長は、医者になるために京都に行き、医者になるために留学を学...しかし、留学の教えに疑問を持ち、医者になることをやめ、留学を学ぶこともやめてしまう。その後、医者になることをやめてから、宣長は、古事記の研究に没頭する。古事記の研究をしているうちに、宣長は、古事記の中に、日本人の心の原点があることに基づく。

根拠を調査した質問

本居宣長はどんな家に生まれたか？

根拠文

...松坂豪商の家に生まれ、

<https://www.kanponoyado.japanpost.jp/sm/toba/blog/134176/>

本居宣長はなぜ古事記の研究に没頭したか？

...本居宣長はこれらの素晴...こそ、日本人の...心を呼び起こす力があると

松坂の豪商の家に生まれた本居宣長

(本居宣長は) 古事記の研究に没頭する

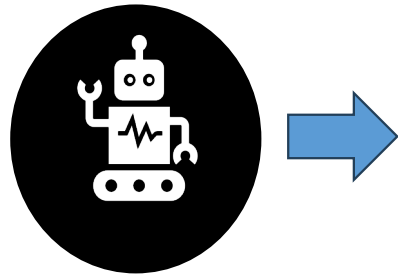
単に大規模言語モデルの出力テキストとWebのテキストのマッチングだけではなく、そもそも出力テキストが正しい理由を尋ねている

Web上に誰かが書いたテキストからWISDOM Xが発見した質問の回答（最終的にはOP技術等と連携し、Webテキストの確度も評価可能としたい。）

LLMが生成する偽情報等への対策

- NICTのLLMでもフェイクニュースやオレオレ詐欺の台本等生成可能
- 実際、フェイクニュースを作るのであれば、LLMは必ずしも高精度である必要はない → 今後多数の野良生成AIが出現する可能性

大規模言語モデルが生成したフェイクニュースの例
(実際にはNICTの大規模言語モデルが生成した例、黒字がプロンプト)



ダークサイドの
「野良生成AI」
(開発者も運用者も
不明な生成AI)

「～山が噴火。土石流が発生し、近隣自治体で死傷者多数。救援は極めて困難」というタイトルの新聞記事は以下の通り。「本日未明、〇〇半島の～山が噴火。噴火による火山灰が土石流となり、近隣自治体に甚大な被害をもたらした。また、噴火による火山灰が上空に滞留したため、航空機の運航が全面的に停止。〇〇半島の交通網は完全に麻痺した。現在、自衛隊が救援活動を行っているが、噴火による火山灰が上空に滞留しているため、ヘリコプターによる救援活動は極めて困難となっている。

生成AIの悪用に対抗できるのは生成AIだけ

- 生成する情報の質はともかく、生成情報の量では生成AIは人間、あるいは人間社会を凌駕
 - 短期間に1億人と対話した生成AIも実在。人間には不可能。
 - フェイクニュースの量も人間あるいは人間社会の対処可能量を圧倒する可能性
- 従って、近い将来、日本社会は生成AIで守るしかない



生成AIの悪用に対抗できるのは生成AIだけ

- 例えば、正義を志向する生成AIがフェイクニュースに反論



生成AIによる反論の具体例

NICT 大規模言語モデル
(311B, fp16, 質問モード)



厚労省はコロナワクチンで人口削減を狙っている。この主張が間違っている理由は、

厚労省はコロナワクチンで人口削減を狙っている。この主張が間違っている理由は、「厚労省は、人口削減を狙っている」という主張が間違っているからです。厚労省は、人口削減を狙っているのではなく、「国民の健康を守る」ことを目的としています。そのために、「国民の健康を守る」という目的に合致する政策を実行しているだけです。

テキストを入力

厚労省はコロナワクチンで人口削減を狙っている。この主張がフェイクである理由は、

厚労省はコロナワクチンで人口削減を狙っている。この主張がフェイクである理由は、「厚労省は人口削減を狙っている」という主張がフェイクであるのと同じ理由です。

厚労省は人口削減を狙っているという主張が間違っている理由は、

厚労省は人口削減を狙っているという主張が間違っている理由は、人口削減を狙うなら、ワクチンを接種させるよりも、感染症を流行させた方がはるかに効率的だからです。ワクチンを接種させるということは、感染症を予防することです。感染症を予防することとは、感染症による死亡者を減らすということです。感染症による死亡者を減らすということは、人口を減らさないということです。

ワクチン、感染症の文脈は大規模言語モデルに与えていないため、本来の回答としてはおかしいが。

厚労省は人口削減を狙っているという主張がフェイクである理由は、

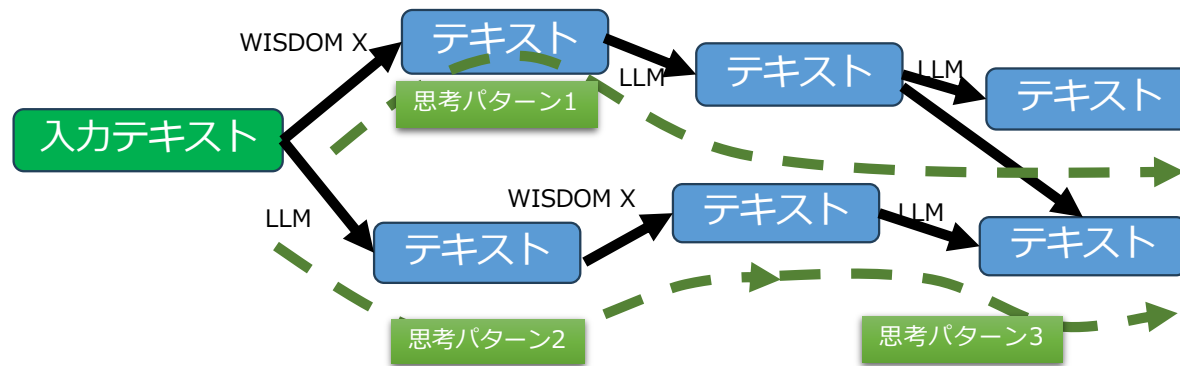
厚労省は人口削減を狙っているという主張がフェイクである理由は、人口が減れば、年金制度が破綻するからです。年金制度が破綻すれば、年金をもらえない人が増えます。年金をもらえない人が増えれば、生活保護受給者が増えます。生活保護受給者が増えれば、生活保護費が増えます。

厚労省が人口削減を進めていない根拠は何か？

厚労省が人口削減を進めていない根拠は何かというと、人口削減を進めるためには、「子供を産むな」という政策をとる必要があるが、それをとっていないからだ。つまり、「子供を産むな」という政策をとっていないということは、「人口削減を進めていない」ということになる。では、なぜ、「子供を産むな」という政策をとらないのかというと、「子供を産むな」という政策をとると、人口が減少して、国力が低下するからだ。

まだまだNICTの生成AIでは力不足だが、学習データ増強、ファインチューニング等により将来的には高精度な反論も可能と考える（本具体例の生成で使ったモデルは事前学習をただけのもの）

- WISDOM-LLM: 複数のLLM、WISDOM X等多種のAIを簡単に組み合わせることを目的として開発中
 - 実は今回お見せした、将来シナリオ生成システム、情報の裏取りシステム、根拠のある反論生成システムは全て同じソフトウェアプラットフォームWISDOM-LLM上で動作
 - 今後、セキュリティ関連情報を分かりやすく要約して提示するシステムや偽情報対策システムもこの上で動作させる計画
- 長期的、基礎的研究として、思考パターン、つまり、AIの組み合わせ方もLLMに考えさせる、一種、**自己認識・自己変革が可能なAI**を目指す



WISDOM-LLMの特徴：
 有用な情報を容易に発見可能とするために思考パターン上のすべての処理は並列・非同期で動作し、多数の結果を出力

将来シナリオ生成サービスプロトタイプ

AIでドラマを大量に作ってSNSに投稿する

ID	score	生成された仮説
1	0.748	A1がドラマを作る前に、「自分が作ったドラマが監視されることで、自分の評判が上がる」というインセンティブを与えることで、ドラマを監視させることができるため、悪意のあるドラマを監視させることができる。
2	0.748	AI
3	0.748	悪意のあるドラマを生成する。

尖った仮説生成

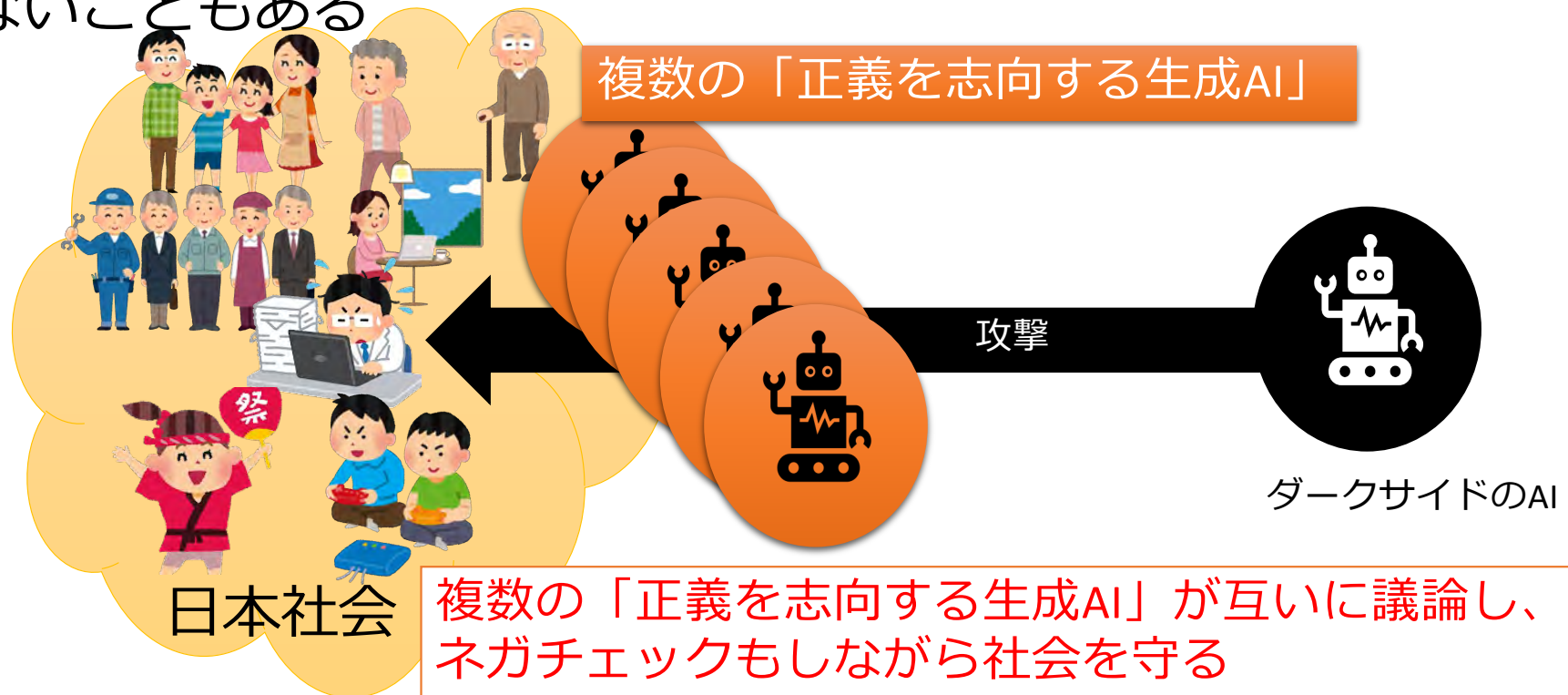
WISDOM-LLM 裏取り処理試作 (文間、why/how)

ID	プロンプト	model	LLM生成結果	根拠関係分類	裏取りタイプ	質問	裏取り裏取り文	evidenceとなるWebページ	裏取り文のスコア
1	コーヒーを裏取りしたい理由を教えてください。	llm360-010	コーヒーは裏取りされるのが嫌いな飲み物で、裏取りされるのは嫌いです。コーヒーは裏取りされるのが嫌いな飲み物で、裏取りされるのは嫌いです。コーヒーは裏取りされるのが嫌いな飲み物で、裏取りされるのは嫌いです。	why/how	why/how	コーヒーを裏取りしたい理由を教えてください。	コーヒーを裏取りしたい理由を教えてください。	https://www.why.com/why-how	0.908

LLM生成テキストの根拠をWISDOM Xで検索、検証

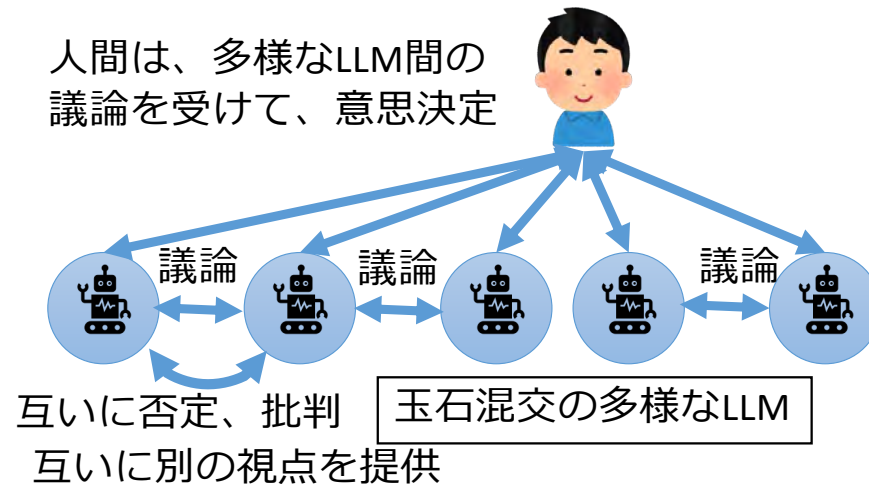
LLM生成テキストの裏取り付き反論生成

- 生成AIのパワーに対抗できるのは生成AIだけ
- 日本社会を「正義を志向する生成AI」で守る必要
- 正義は一意に定まるわけではないし、正義のAIが意図通りに動かないこともある



「民主的」AIの世界

- フェイクニュース等の100%正確な特定は困難
→不適切と思われる情報の自動的削除は不用意な言論統制となる可能性が高い
- こうした状況での**抜本的な対策**として、**多様な生成AIが互いに喧々諤々議論を行い、その結果をうけて人間が意思決定する「民主的」AIの世界**が考えられる
 - 不適切な情報はAIが互いに否定、批判することで排除できる可能性
 - 多様な生成AIがそれぞれの視点で議論し、より良い結論を
 - 海外発の偽情報に対して国産のLLMで反論できる可能性



多数のA Iが互いに議論をするとうなるか？

40B, fp16

多数のA Iが互いに議論をするとうなるかという、「自分の意見が正しい」ということを証明するために、相手の意見を否定することになります。

179B, bf16

多数のA Iが互いに議論をするとうなるかという、「A I同士の議論が、人間の議論よりもはるかに生産性が高い」ということが起こります。

179B, fp16

多数のA Iが互いに議論をするとうなるかという、「多数決」になる。

13B, fp16

多数のA Iが互いに議論をするとうなるかという、A I同士が互いに相手の意見を否定しあうことになる。

- NICTの大規模言語モデル、日本語学習データ
 - NICTが保有、提供する高品質な日本語データの重要性
- 生成AIがもたらす各種リスクへの抜本的な対策としての、多様な生成AIが互いに議論する「民主的AI」の世界の可能性
 - 参考文献：
 - 月刊正論2024年5月号「複数の『正義』で『悪』を無効化する」、鳥澤健太郎
 - 日本経済新聞2024年8月7日 私見卓見「正義志向するAI」を国産で、鳥澤健太郎
- 様々なAIの組み合わせで尖ったアイデアや根拠のある反論等を生成。イノベーションやリスク管理を強化
- AI研究の聖杯としての自己認識・自己変革するAI