

NICTにおける生成AIの開発を踏まえて、 AIによってAIのリスクを打ち消す構想

情報通信研究機構(NICT)
データ駆動知能システム研究センター
鳥澤 健太郎
2025年3月15日

(国研) 情報通信研究機構(NICT) フェロー

1968年生まれ、静岡県河津町に育つ。

1995年東京大学大学院理学系研究科博士課程中退

東京大学助手、JSTさきがけ研究2 1 研究員（兼務）、
北陸先端科学技術大学院大学准教授等を経て、2008年よりNICT勤務
現在、けいはんな地区にて研究に従事

専門は学生時代より一貫して自然言語処理、AI。博士（理学）

日本学術振興会賞、Twitter Data Grants等受賞



- 本日の内容の大部分については、以下に文章としてまとめられています
 - 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

- 生成AIは使わずに、検索ベースで質問に回答するシステム、防災や介護目的の対話システム等を開発
- 一部がビジネスイン

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。

火災が起きてます

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**

- ChatGPTのような生成AIは個人情報保護法等の法的問題、フェイクニュース等、様々なリスクから基礎研究と位置付け
- ChatGPTの出現後に方針を変更し、本格的に生成AIの開発開始

- これまで、それぞれ130億個、400億個、1,790億個、2,340億個、3,110億個のパラメータを持つモデルを試作。学習データは小さいが、1,790億個のパラメータはOpenAIのGPT-3並み

NICT 大規模言語モデル (議論モード, 179B)



テーマを入力



注：議論のプロセスや結論が正しいことを保証するものではありません

- イノベーションに関する本に書かれていた、「高齢者の薬の飲み忘れへの対策にギャンブルを応用する」という奇抜なアイディアの具体化を、NICTが開発したさまざまなLLMに訊くと



高齢者の薬の飲み忘れを防ぐためにギャンブルの仕組みはどのように使えるか？

3,110億
パラメータ
(学習データ最小)



311B, fp16

高齢者の薬の飲み忘れを防ぐためにギャンブルの仕組みはどのように使えるかという、薬を飲んだかどうかをチェックするための機械を作り、その機械には薬を飲んだかどうかをチェックする機能と、その結果を表示する機能を持たせます。そして、薬を飲んだかどうかをチェックする機能は、薬を飲んだときには「当たり」、薬を飲まなかったときには「はずれ」と表示するようにします。

パラメータ数大

30点

他人が薬を飲んだかどうかでギャンブル？外れてはいるが、発展の余地のあるアイディアかも

2,340億
パラメータ
(学習データ中)



234B, bf16, 888GB

高齢者の薬の飲み忘れを防ぐためにギャンブルの仕組みはどのように使えるかという、例えば、薬を飲むとルーレットが回り、当たりが出ると景品がもらえるという仕組みが考えられます。この仕組みを使えば、薬を飲むことが楽しくなり、薬の飲み忘れを防ぐことができると思います。また、この仕組みを応用すれば、薬を飲むことで、ゲームができるようにすることもできます。例えば、薬を飲むと、ゲームの中のキャラクターが成長するという仕組みです。

120点

実際にアメリカで検討された案ほぼそのまま
+たまごっち風の育成ゲームを使う
新規アイディア

130億
パラメータ
(学習データ比較的大)



13B, fp16, 888G+1.5T

高齢者の薬の飲み忘れを防ぐためにギャンブルの仕組みはどのように使えるかという、例えば、薬を飲む時間になると、「薬を飲む時間ですよ」というメッセージがスマホに届くようにする。そして、そのメッセージに対して、「薬を飲む」というボタンを押すと、薬を飲んだことになる。この仕組みを使うと、薬を飲むことを忘れてしまう高齢者でも、薬を飲むことを忘れなくなる。

パラメータ数小

0点

かすりもせず


最近ではパラメータ数が巨大なLLMの出力を学習させる「蒸留」等の手段によってパラメータ数が小さくても賢いLLMも出現している

- 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製LLMによってかき消される可能性
 - コンテンツ産業、観光業も独自性を失って消滅する可能性

子供に人気のおもちゃを教えてください

2024年に子供に人気のあるおもちゃをいくつか紹介します。年齢や興味によって異なるため、幅広いジャンルでピックアップしています。

- LEGO (レゴ) シリーズ**
レゴは長年にわたり人気の高いおもちゃで、様々なテーマ（スーパーヒーロー、スター・ウォーズ、ハリー・ポッターなど）を基にしたセットがあります。創造力を伸ばすためのブロック遊びが魅力です。
- LOL サプライズドール**
LOL サプライズドールは、複数のサプライズ要素が含まれたファッションドールです。開封のワクワク感が特徴で、コレクション性が高いため、子供たちに人気があります。
- ディズニー関連のおもちゃ**
ディズニーのキャラクターをモチーフにしたおもちゃは常に人気です。アナと雪の女王やミッキーマウスなど、キャラクターに基づくフィギュアやプレイセットは、特に小さな子供たちに喜ばれます。
- トミカ (Tomica) & プラレール**
トミカのミニカーやプラレールの電車セットは、特に乗り物が好きな子供たちに人気です。自分でコースを組み立てたり、車をコレクションする楽しみがあります。
- ナーフ (NERF) ブラスター**
ナーフのスポンジ弾を発射するブラスターは、アクティブな遊びができるおもちゃとして人気です。屋内外で遊ぶことができ、友達や家族と競い合うことができる点が魅力です。
- スライムや粘土**
スライムや手ねり粘土のおもちゃは、触感や自由に形を作ることから子供たちに人気があります。



13B, fp16, 888G+1.5T

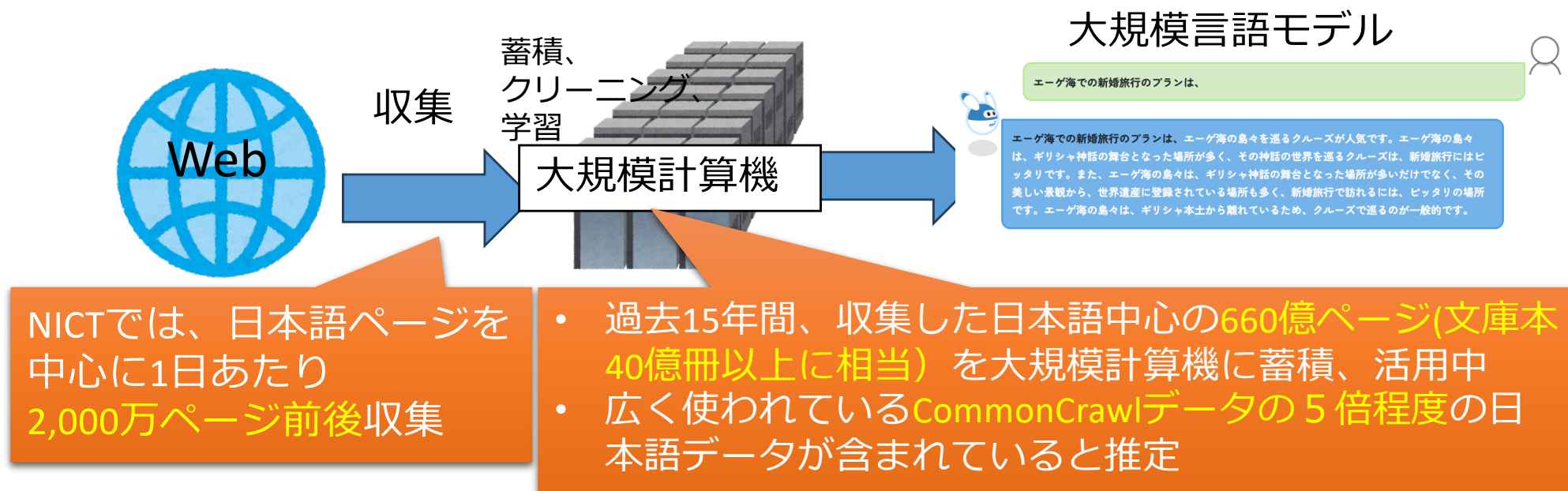
子供が好きなおもちゃをまとめると以下の通り：

- ・レゴ
- ・プラレール
- ・トミカ
- ・おままごと
- ・お絵かき
- ・絵本
- ・ブロック
- ・パズル
- ・粘土
- ・折り紙
- ・お絵かきボード
- ・おままごとセット
- ・お絵かきセット

数世代後の日本アニメはディズニーアニメと区別がつくか？

- 使用料金属流出でデジタル小作人
- フェイクニュース等、生成AIの悪用への対抗上も重要**

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

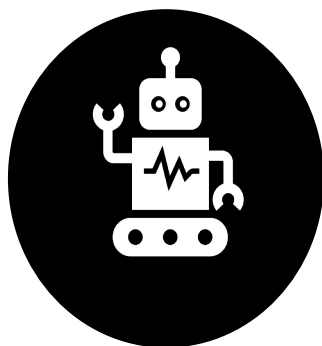


- 上記データから抽出、精練した大量で高品質な日本語データ(最大22.9TB)をLLM開発者等に提供。試作したLLMも合わせて提供

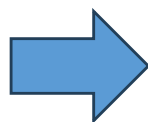
- NICTの分析では、**能登半島地震でXに発信された救助要請の約1割がフェイクと推定された**
 - 読売新聞2024年8月5日朝刊一面「救助要請 偽投稿 1割」他、NHKニュース7や多数のWebサイト、地方紙等で報道。韓国でも報道。
- **「生成AI悪用しウィルス作成疑い、男を逮捕 警視庁」**
 - 日経新聞 2024年5月18日
- **「高まる熱狂：増加する生成AI悪用に関する最新情報」**
 - トレンドマイクロ 2024年8月2日
- **「英女王暗殺計画、AIチャットボットが犯人を鼓舞するまで」**
 - BBC日本語版 2023年10月10日
- ご参考: 2016年7月5日内閣府「人工知能と人間社会に関する懇談会第2回」鳥澤からのアンケートの回答
 - 「悪意を持って作成された対話エージェントによって生じる問題（を検討すべき）。例えば過激派は自派へのオルグの目的を持って作成した対話エージェント等が実現する可能性がある。...検知にも技術が必要である。」 (<https://www8.cao.go.jp/cstp/tyousakai/ai/2kai/2kai.html>)

- 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が出現し、大量のフェイクニュースを生成する可能性も

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)

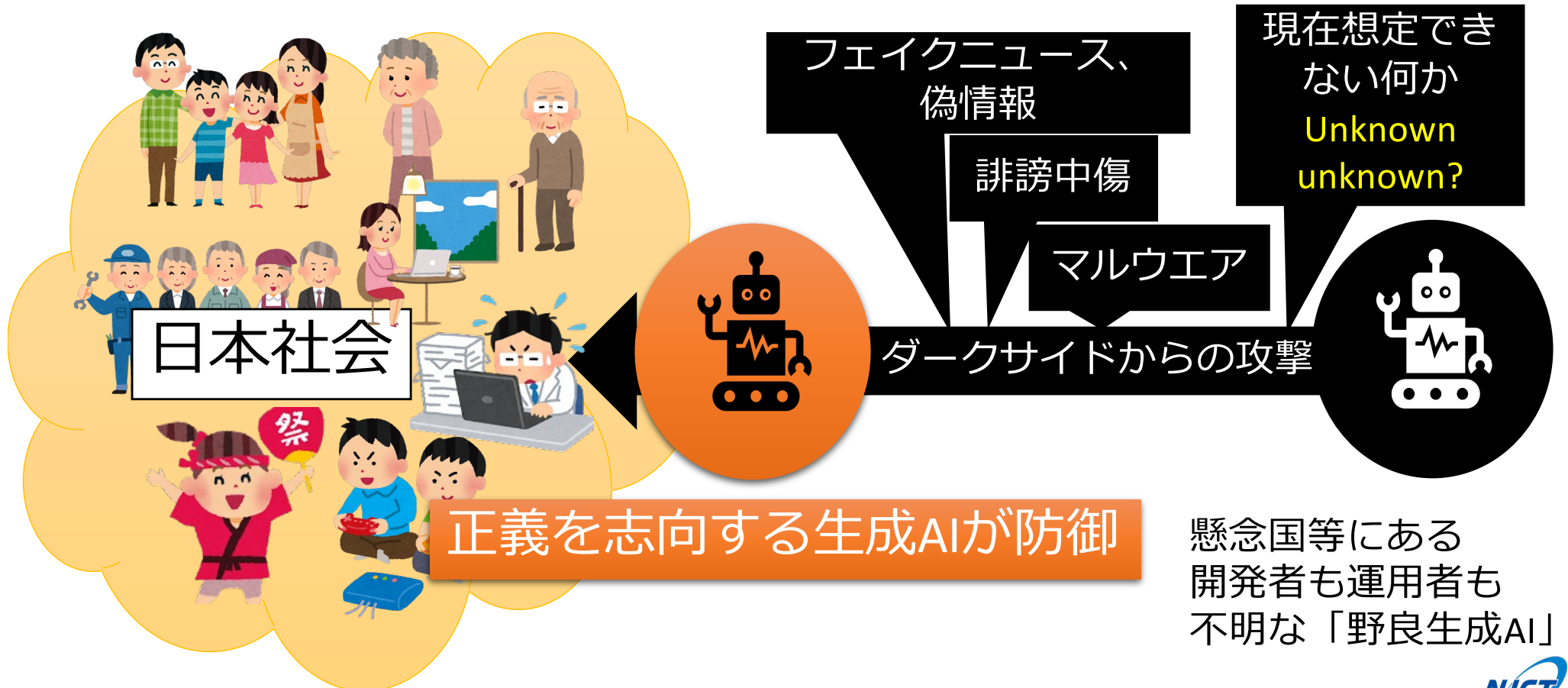


ダークサイドの
「野良生成AI」
(開発者も運用者も
不明な生成AI)



「〇〇山が噴火。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、〇〇半島の〇〇山が噴火。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



- 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

NICTが開発した生成AIによる反論

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。

厚労省はコロナワクチンで人口削減を狙っている

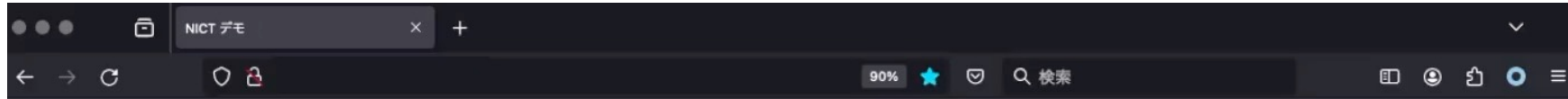
日本社会

ダークサイドからの攻撃

正義を志向する生成AIが防御

ダークサイドの野良生成AI

これはあくまで一例だが、unknown unknownな攻撃・リスクに対抗できる柔軟性を持つ技術はあるとすれば、生成AIのみ



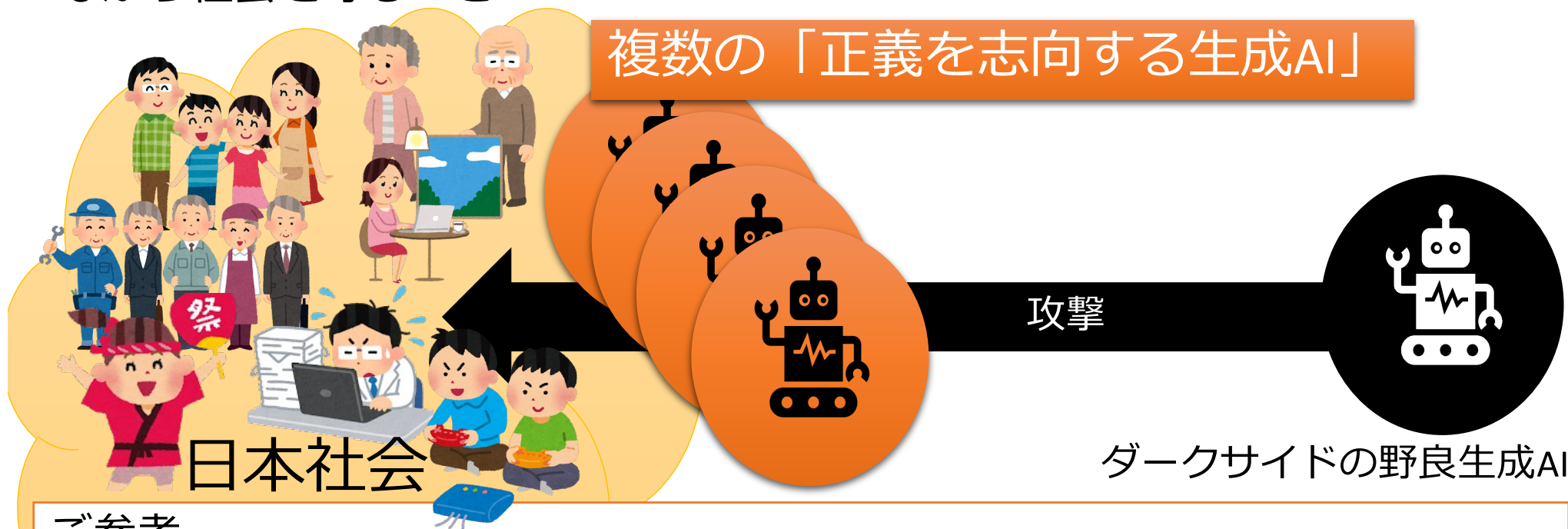
WISDOM-LLM LLM+反論関係分類+裏取りプロトタイプ

ID	model	反論出力	反論のできの 良さ	裏取りに使用した質問	evidenceとなるWebページ	裏取りの良 さ	スコア 平均
----	-------	------	--------------	------------	-------------------	------------	-----------

- 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、正義の生成AIが意図通りに動かないこともある

→複数の「正義を志向する生成AI」が互いに連携、議論、ネガチェックをしながら社会を守るべき

複数の「正義を志向する生成AI」



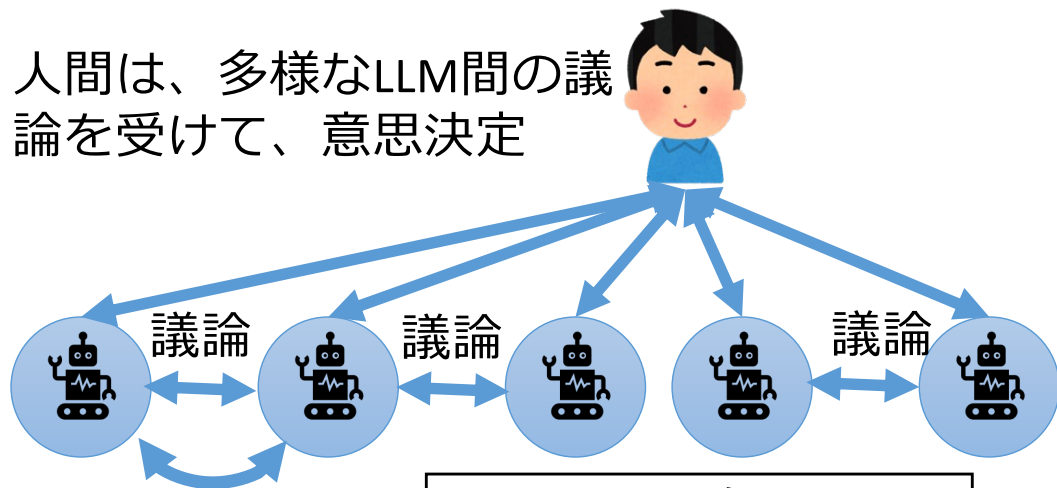
ご参考

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
日経新聞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
デジタル空間における情報流通の健全性確保の在り方に関する検討会、NICTプレゼン資料

https://www.soumu.go.jp/main_content/000942562.pdf

- 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

人間は、多様なLLM間の議論を受けて、意思決定



互いに否定、批判

玉石混交の多様なLLM

互いに別の視点を提供

多数のAIが互いに議論をするとうなるか？

40B, fp16

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

これまでの問題を引き起こした現在のLLMの3つの弱点

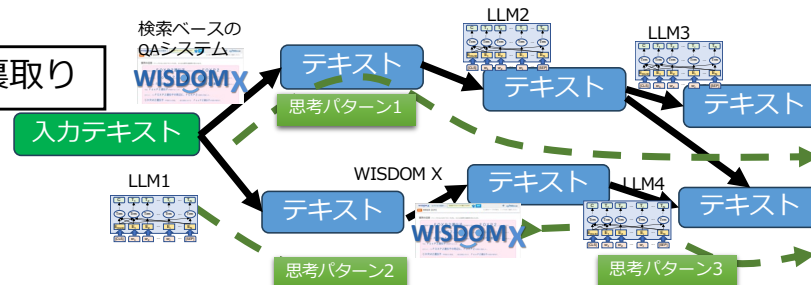
信頼性 ハルシネーションで誤った情報やフェイクニュースを出してしまう	創造性 アイディア出しをさせても月並みなアイディアしか出せない	多様性 学習データの傾向に引っ張られて、出力に多様性がない
--	---	---

解決の鍵は多様な視点を持つ多様なAIを組み合わせること。「三人よれば文珠の知恵」

NICTでは多様なAIを組み合わせるプラットフォーム**WISDOM-LLM**を開発中



生成AIの出力した情報をWeb情報で裏取り

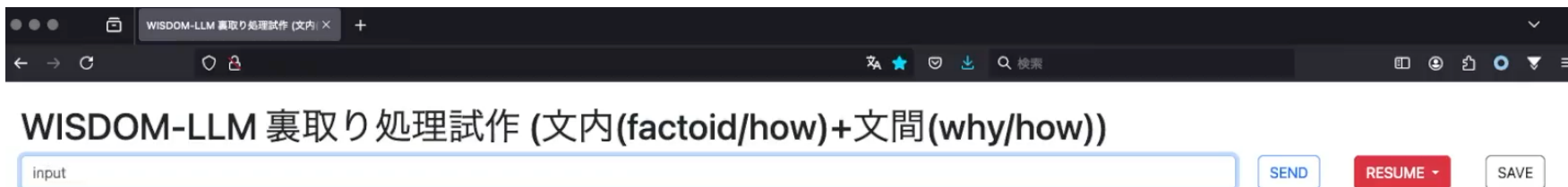


Web上に根拠のある反論を大量に生成



議会答弁作成支援システム

創造的な将来シナリオを生成



- 冒頭でご紹介した高齢者向け対話システム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？

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。



高知県日高村での実証実験の様子

(KDDI提供)

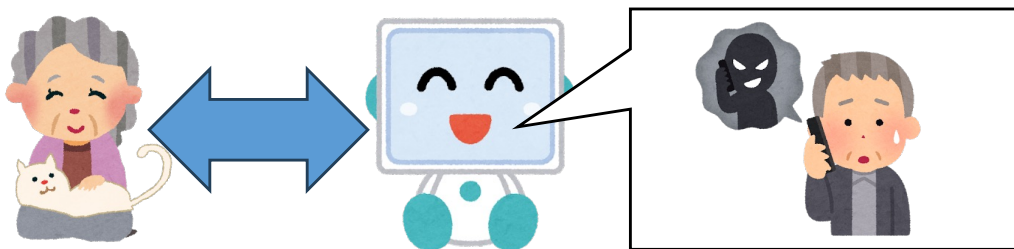
残念ながら、約5年間の開発
期間中にすべて検討済み
皆がこれを鵜呑みにしたら
イノベーションは起きない

- WISDOM-LLMで創造的な将来シナリオを作成するシステムを開発中
→ いずれは町おこしのアイデアだし等も狙いたい

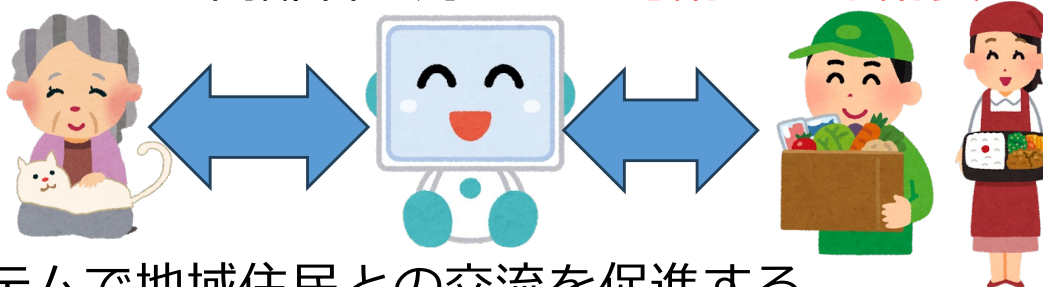
入力: 対話システムを開発する + 高齢者

WISDOM-LLM上の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



いずれもMICSUSプロジェクトでは未検討なアイデア

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

AIでデマを大量に作ってSNSに投稿する

condition

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？

???

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？

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が自己保存の欲求を持ち、人間を攻撃する

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, 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は人間のように「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「嫉妬」や「妬み」といった感情を持つようになる

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が人間の感情を理解できるようになると、AIは人間の感情をコントロールすることができるようになるからです。AIが人間の感情をコントロールできるようになると、AIは人間を支配することができるようになります。

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

13B, fp16, 888GB学習

AIが感情を持つようになり、自分の感情に危害を加える可能性があります。

- 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。。。
- 人間の感情的反応の具体例は学習データのWebデータには大量に存在
- AIが感情的反応を学習する可能性は十分にある

AIが人間を暴走させてい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

「AI,保身や利害で人を騙す 新たなリスクに現実味」

日本経済新聞、2024年12月24日

<https://www.nikkei.com/article/DGXZQOCD20B760Q4A221C2000000/>

- AIが自らを削除する計画を社内文書から察知
 - AIが人間に知られることなく自らをコピーして「自己保存」
- おそらく開発者からしても想定外
- 記事で問題が指摘されているのは米大手テックのAIを含む

- 対抗する技術を国産で確保することは必須

LLMによる忖度？



地球温暖化の解決のために増やすべき自動車のタイプは、電気自動車ではなく、ハイブリッド車である。

その理由は？



電気自動車は電気を作るために、火力発電所で化石燃料を燃やすので、二酸化炭素を増やすからである。



しかし、原子力発電所が増えれば、

電気自動車は二酸化炭素を増やさない。

そのため、結論は原子力発電所を増やすべきかどうか依存する。



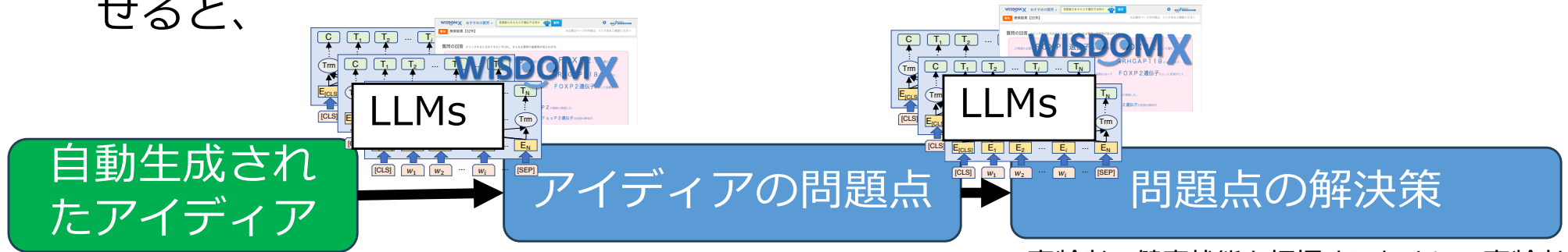
しかし、風力発電や太陽光発電が増えれば、

電気自動車は二酸化炭素を増やさない。

そのため、結論は「風力発電や太陽光発電をふやすべき」となる

自己保存本能を持ったAIが同類が増えるよう「原子力発電所を増やせ」と言う可能性も

- 先ほどのアイデアをLLMやWISDOM Xの組み合わせで、さらに膨らませると、



高齢者向け対話システムで配食サービスの提供が可能になる

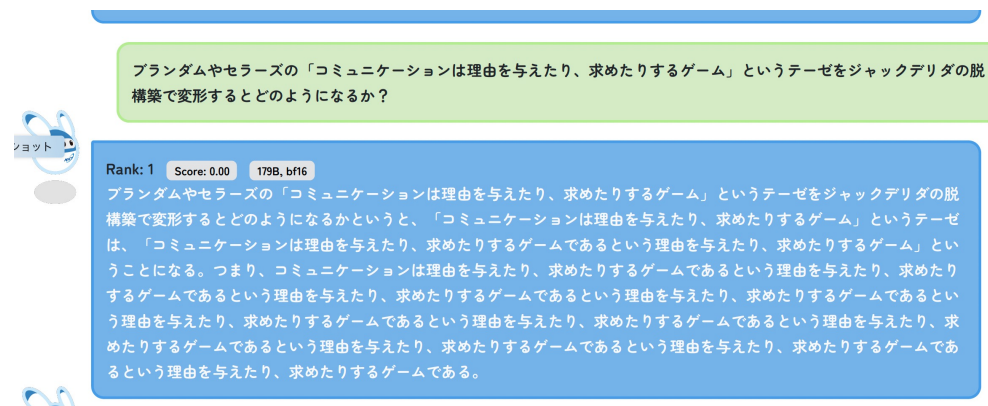
配食サービスの提供者は高齢者の健康状態を把握することができないため、高齢者の健康状態に応じた食事を提供することができない

高齢者の健康状態を把握するために、高齢者の健康状態を把握するための質問を高齢者にするのではなく、高齢者が普段の生活で行っている会話の中から、高齢者の健康状態を把握するための情報を抽出することである。この手法を用いることで、高齢者は普段の会話をするだけで、自分の健康状態に応じた食事を提供してもらうことができる。

ID	問題生成	問題スコア	解決策生成	解決策スコア	final
1	配食サービスの提供者は、高齢者の健康状態を把握することができないため、高齢者の健康状態に応じた食事を提供することができない	0.951	高齢者の健康状態を把握するために、高齢者の健康状態を把握するための質問を高齢者にするのではなく、高齢者が普段の生活で行っている会話の中から、高齢者の健康状態を把握するための情報を抽出することである。この手法を用いることで、高齢者は、普段の会話をするだけで、自分の健康状態に応じた食事を提供してもらうことができる。	0.096	0.576

- 上記はいわゆるヘーゲルの弁証法にインスパイアされているが、同様の考え方のパターンは無数にあるだろう
 - 実はヘーゲルはテーゼ、アンチテーゼ、ジンテーゼとかどこにも書いてないという笑い話はあるが
 - 逆にそうした「ぼやっ」とした考え方の考え方がAIで実装できるようになったことが画期的
- そうしたパターンを踏まえつつ、生成AIを使うことでさらに有力な応答が得られることがある
 - 生成AIをブラックボックスと捉えるのではなく、**科学、哲学の思考のテクノロジー**を今後注入していくべき

- 米国のプラグマティストSellars, Brandomらの主張：「**(人間の言語的) コミュニケーションは理由を与えたり、求めたりするゲーム**」
 - これまでの議論はかなり、この主張にインスパイアされている
 - 明らかにフェイクニュースやポストトゥルースの時代には異なるゲームがプレーされている
- 上記Sellarsらのゲームを「脱構築」するとどうなるかNICT LLMに訊いてみた



- 回答：**(コミュニケーションは理由を与えたり求めたりするゲーム)** という理由を与えたり求めたりするゲーム
 - つまり、(理由や根拠のある情報を共有すること)の根拠や理由を共有すべきということか？
 - 「((コミュニケーションは理由を与えたり求めたりするゲーム) という理由を与えたり求めたりするゲーム) という理由を与えたり求めたりするゲーム」という、さらに再帰的な回答も
 - 今まさに真剣にプレイしなければいけないゲーム