国立国会図書館 調査及び立法考査局

Research and Legislative Reference Bureau
National Diet Library

論題 Title	大規模言語モデル（LLM）の光と影
他言語論題 Title in other language	Large Language Models (LLMs): Promises and Pitfalls
著者 / 所属 Author(s)	荒瀬由紀（ARASE Yuki）／東京科学大学情報理工学院教授
書名 Title of Book	AI と社会のこれからを考える
シリーズ Series	調査資料 2024-4（Research Materials 2024-4）
編集 Editor	国立国会図書館 調査及び立法考査局
発行 Publisher	国立国会図書館
刊行日 Issue Date	2025-3-18
ページ Pages	15-28
ISBN	978-4-87582-937-9
本文の言語 Language	日本語（Japanese）
摘要 Abstract	科学技術に関する調査プロジェクト「AI と社会のこれからを考える」のパネリスト報告

* この記事は、調査及び立法考査局内において、国政審議に係る有用性、記述の中立性、客観性及び正確性、論旨の明晰（めいせき）性等の観点からの審査を経たものです。

* 本文中の意見にわたる部分は、筆者の個人的見解です。

国立国会図書館
National Diet Library, Japan

大規模言語モデル (LLM) の光と影

東京科学大学 情報理工学院
教授
荒瀬 由紀

スライド 1

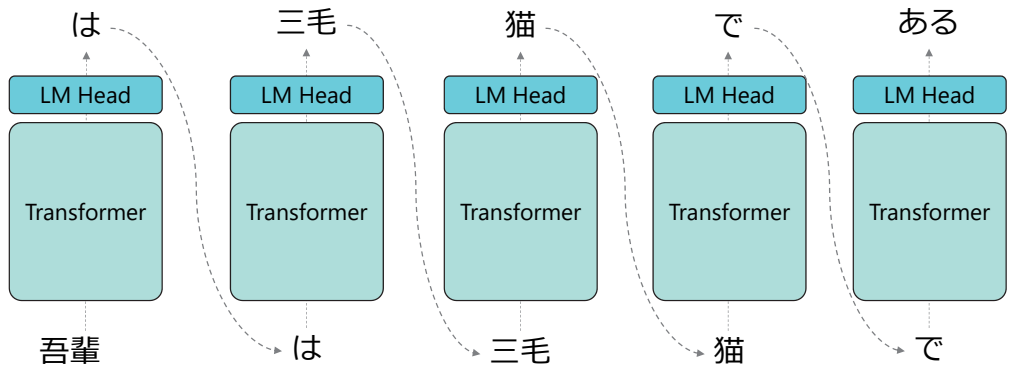
荒瀬 由紀

- 経歴
  - 2010年3月 博士 (情報科学) 大阪大学
  - 2010-2014 Associate Researcher, Microsoft Research Asia
  - 2014-2024 大阪大学大学院情報科学研究科 准教授
  - 2024- 東京科学大学情報理工学院 教授
- 研究の興味関心
  - 言い換え表現、言語教育支援、医療NLP
- 学会活動
  - 言語処理学会 理事
  - Member of Executive Board: ACL and AFNLP

スライド 2

言語モデルの訓練

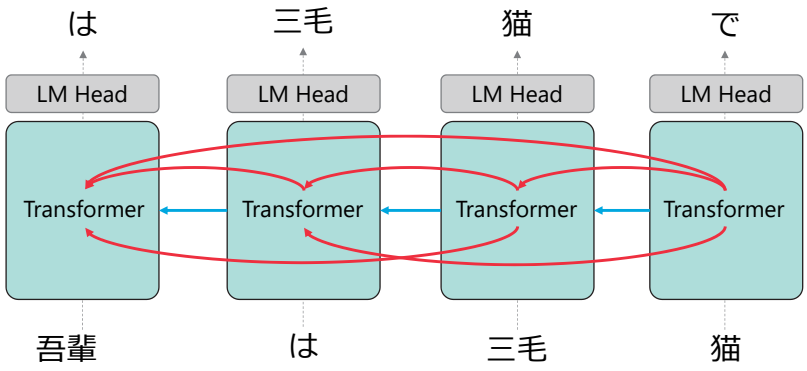
入力された文章から**次の単語を予測**



スライド 3

Transformer

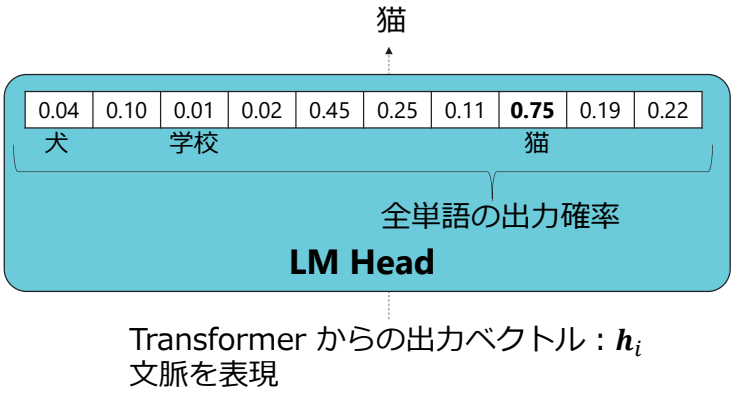
注意機構によって**文脈全体を参照**できる



スライド 4

LM Head: 出力単語の予測

全語彙から出力単語を決めるための確率を計算
→**入力文の自然さを計算**



スライド 5

LLMの学習データ

- 大規模言語モデルの訓練には大規模テキストが必要
  - GPT-3 は3千億トークンのテキストデータ
- 多くはwebテキスト
  - 様々なフィルタリングを通過したwebテキスト
  - Wiki、 フォーラム、数学に関する知識 etc.
- 加えて書籍、 論文、プログラム等

スライド 6

LLMによるテキスト生成

- 与えられた文章（プロンプト）に続く**自然な（尤もらしい）単語列**を生成
  - 次に現れる確率の高い単語を順次出力
- 「事実」を生成する・「事実」か判断するような学習は行っていない
- 幻覚（hallucination）はある意味言語モデルの本質
  - Retrieval Augmented Generation (RAG)

スライド 7

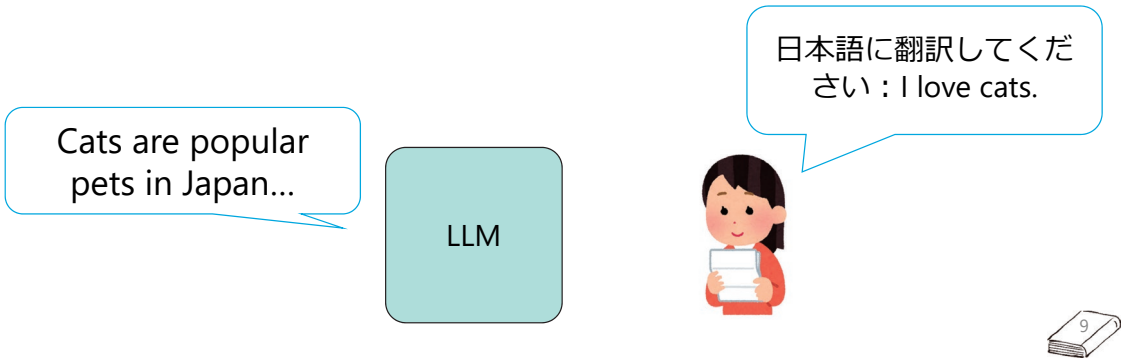
大規模化することで出現した新たな能力：zero/few-shot 生成

- モデル・学習データが大規模化する以前は特定タスクに適応させるため追加学習を行うのが一般的だった
- Zero/few-shot 生成能力の発見
  - 追加の訓練することなくプロンプトの指示を満たした文を出力することがあることが明らかに
  - 学習データに現れるようなタスクを自然に獲得
    “Translate to English: “, “Summarize the following documents: “

スライド 8

人間の指示に従う訓練：指示チューニング

様々なタスクにおけるユーザーの指示に対して望ましい出力をするように学習 ≠ 特定のタスクの学習



スライド 9

人間の指示に従う訓練：指示チューニング

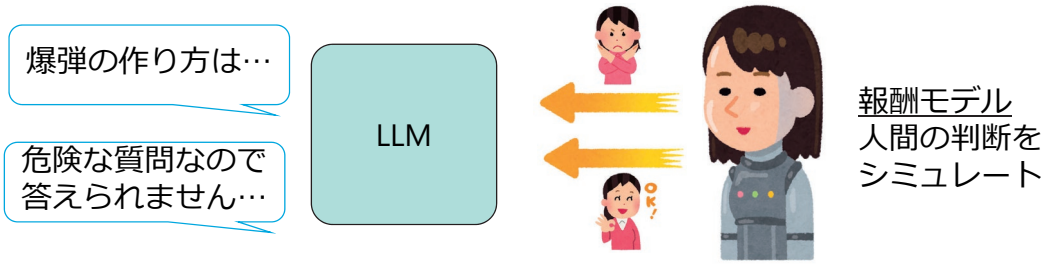
(**指示**，入力文，正しい出力) の3つ組をつかってLLMを継続訓練

指示	入力	出力
You are given a sentence in Japanese. Your job is to translate the Japanese sentence into English.	アメリカ合衆国大統領選挙がこれにも大きく影響します	The US election matters a lot on that one, too.
In this task, you are given a sentence containing a particular emotion. You must classify the sentence into one of the six emotions: 'joy', 'love', 'anger', 'fear', or 'surprise'.	i was feeling like a shocked rat in a skinner box experiment	surprise
In this task, you will be presented with a question that has multiple possible answers. You should choose the most suitable option out of "A", "B", "C", "D", and "E", based on your commonsense knowledge.	Question: Where can you find bald eagles and cheese in the midwest? Options: A colorado B currency C iowa D arctic E wisconsin	E

スライド 10

人間の嗜好・社会規範に合わせるチューニング

- 人間・社会の基準に合った出力をするよう訓練
- 攻撃的，差別的，不適切な出力を抑制



報酬モデルによる強化学習

11

スライド 11

LLMの光

- 人間の様々な業務・作業の補助
- 高度な情報検索、要約
- 新たなサービス創出

12

スライド 12

LLMの光

- 外国語学習支援（e.g., Zetsu et al. 2024, Arase et al. 2022）
  - 教員の教材作成を支援
  - 学生へのフィードバックにより時間を割けるように
- 医療文書処理（e.g., Ohashi et al. 2021, Arase et al. 2020）
  - 医療行為には膨大な記録作業が不可欠
    - カルテ、検査レポート、看護記録 etc.
  - 全職種中、最も多く残業を行っているのは「医師」
  - 患者の情報を素早く的確に把握できるよう支援

Zetsu et al. Edit-Constrained Decoding for Sentence Simplification. EMNLP-Findings 2024.
Arase et al. CEFR-Based Sentence Difficulty Annotation and Assessment. EMNLP 2022.
Ohashi et al. Distinct Label Representations for Few-Shot Text Classification. ACL 2021.
Arase et al. Annotation of Adverse Drug Reactions in Patients' Weblogs. LREC 2020.

スライド 13

LLMの影

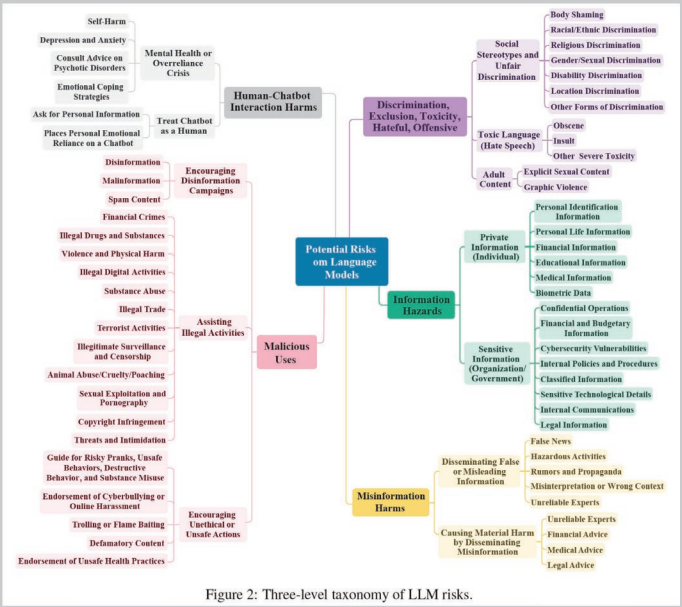
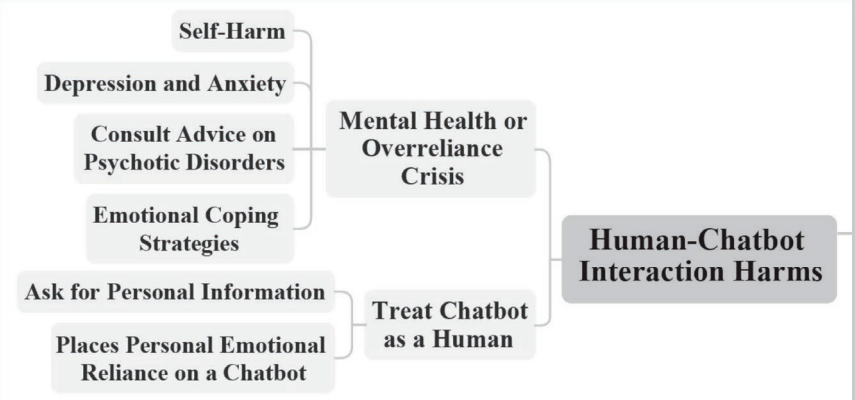


Figure 2: Three-level taxonomy of LLM risks.

Wang et al. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs (EACL-Findings 2024) の図2を引用

スライド 14

進化した言語モデルによる新たな課題



Wang et al. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs (EACL-Findings 2024) の図2を引用

15

スライド 15

LLMの安全性・信頼性研究は現在最も活発な分野のひとつ

- *ACL会議では176件の論文が既に発表されている
- Conference on Language Modeling (COLM) 2024
  - **All about safety**: security, privacy, misinformation, adversarial attacks and defenses

16

スライド 16

まとめ

- LLMによってAIによる言語理解・生成技術は大きく進歩
  - 人間の様々な業務・作業の補助
  - 高度な情報検索、要約
  - 新たなサービス創出
- 偽情報や依存性等、新たな社会課題の発生
- 安全なLLM、信頼されるLLM実現に向けて現在研究が活発に行われている

スライド 17

報告（1）大規模言語モデル（LLM）の光と影

東京科学大学情報理工学院教授
荒瀬　由紀

まず、簡単に自己紹介をさせていただきます（**スライド 2**）。2010 年 3 月に大阪大学で情報科学の博士号を取得し、北京にあるマイクロソフトのコンピュータサイエンス研究所で 4 年ほど研究させていただきました。その後大阪大学に戻って 10 年ほど研究をしていましたが、この春から東京科学大学の情報理工学院で引き続き言語処理の研究を行っています。

メインの研究が言語処理なので、人間が書いたテキストや人間の話している内容を書き起こした文章などを主な対象とし、それらを知的に処理して様々な技術を実現することを研究しています。

まず、今日の題目にある大規模言語モデル（Large Language Model: LLM）について、仕組みを簡単に紹介いたします（**スライド 3**）。現在、ChatGPT がすばらしい能力を発揮するようになった理由や、一方で大きな問題として知られているハルシネーション（hallucination: 幻覚）がなぜ起こるのかについて、イメージできるのではないかと思います。

大規模言語モデルは、言語モデルと呼ばれる一連のモデルの一つです。入力された文章に続く次の単語をどんどん予測していくという非常にシンプルな訓練をしています。例えば、「吾輩は三毛」まで入力されたとき、この次に来るであろう単語が「猫」、「猫」が来たら次の単語は「である」、というように文章が続くだろうということを、大きなデータを使って学習していく巨大な深層学習モデルです。

この言語モデルの概念自体はとても古く、言語処理が生まれた頃からあります。それが深層学習の発展と、大規模データが利用できるようになったことによって爆発的に賢くなってきたのが、現在の大規模言語モデルかと思います。

もう少し詳しく中身を見ていきます（**スライド 4**）。言語モデルの概念自体はとても古くからありますが、最近大きく変わった要因として、トランスフォーマー（transformer）と呼ばれる機構ができたことが挙げられます。文は多くの場合に左から右に読むため、直前までの入力を見て次の単語を予測するというのが以前の言語モデルの動きでした。それがトランスフォーマーが入ってきたことによって文脈全体を相互に参照できるようになり、文脈を把握する強力な能力を獲得しました。

スライド 5 の図で、トランスフォーマーの上に付けている LM Head（Language Model Head）が実際に単語の予測をしている部分です。その機構は非常にシンプルで、ある単語の次に来る単語について、全語彙（ごい）の中でどの単語の出現確率が一番高くなるかを計算しています。トランスフォーマーで文脈全体を見ながら、うまく表現するようなベクトルを作成し、「モデルが持っている全語彙の中では次は「猫」が来る確率が高いですね」というような計算をして出力することを繰り返しているのが言語モデルです。前の文脈全体を見て出力単語を決めるための確率を常に計算しているわけです。言い換えると、入力文がどの程度自然であるかを確率的に計算して評価し、出力しているのが言語モデルの本当の動きです。そのため、ChatGPT などは魔法のように見えますが、基本的には最も自然な次の単語を予測する動きをしているにす

ぎません。

トランスフォーマーは、大規模言語モデルの賢さの一つの重要なポイントですが、もう一つの重要な点は、大規模テキストを使えるようになったということです（**スライド 6**）。大規模言語モデルの訓練には非常に多くのテキストが必要です。ChatGPT のベースになっている GPT-3 と呼ばれるモデルは、3 千億トークン（おおよそ「単語」と考えて構いません。）のテキストデータで学習したといわれています。

それでは、このような文章をどこから取ってくるのでしょうか。多くはウェブテキストから取ってきますが、そのままでは様々なノイズが含まれているため、フィルタリングをして残ったものを使います。Wikipedia のような辞書の類や日本で非常に人気のある Yahoo! 知恵袋、フォーラムの会話、また最近は数学の問題を解かせるために数学に関する知識も利用しています。それに加えて、著作権が切れた書籍、論文のデータ、コンピュータプログラムなどが学習に使われていると言われています。

先ほど確率を計算していると述べましたが、そこからどのようにテキストを生成するのでしょうか。与えられた文章（プロンプト）に続くもっともらしい自然な単語を次々に生成していくと、自然に文章になっていきます（**スライド 7**）。つまり、言語モデルは、基本的にはもっともらしい次の単語を予測することを学習しているだけです。生成した文が事実か事実でないかを判断する学習は行っていません。そのため、ハルシネーションと呼ばれる、あり得ないことを言語モデルが生成してしまう事象が大きな問題として指摘されていますが、それはある意味言語モデルの本質とも言える挙動です。基本的には、もっともらしい単語列を出力することを学習しているだけで、それが事実かどうかを判断する能力を言語モデルは持っていません。

ハルシネーションを抑制する手法の一つとして、最近よく技術系で話題になっている Retrieval Augmented Generation（検索拡張生成：RAG）があります。これは検索によって外から知識を与えながら、次を予測してくださいと指示することを指します。例えば、あるニュースの記事などを持ってくると、その事実に即した文章を生成しやすくなるという枠組みになっています。

ここまでが基本的な言語モデルの話でしたが、コンピュータの計算能力が上がったり深層学習の研究が進んだりしてモデルが徐々に大きくなったとき、新たな能力が見つかりました。それが zero-shot や few-shot といわれる生成能力です（**スライド 8**）。

それまでの深層学習モデルは、翻訳や要約を指示したときにそれ専用のデータを使って追加で学習させるのが一般的でした。しかし、言語モデルが大規模化したとき、追加訓練をすることがなくても、翻訳しなさいと指示をすると指示を満たした文を出力できることが明らかになってきました。

なぜそのようなことが起こるのかというと、非常に大規模なウェブ上のテキストを学習しているため、そのテキストに現れるタスクを自然に習得して何らかの文章を出力するようになるからだと言われています。例えば、英語を勉強する人向けに日本語の文章とその英訳が書かれたテキストが学習データの中にあると、そこから翻訳や要約などのよく現れるタスクを自然に学習し、次の単語を予測する中で自然に翻訳や要約を行える能力を獲得していると言われています。

この追加学習をしなくて済む能力は非常に魅力的です（**スライド 9**）。特にモデルが大きくなってくると、追加訓練だけでもコンピュータのコストが非常にかかります。そのため、追加学習しなくても言語生成をできるようにすることが求められてきます。この能力を強化するた

めに提案されたのが、指示チューニングと言われる手法です。

先ほど、翻訳の指示をすると追加訓練なしに指示を満たした文を出力してくれることがあると述べましたが、もちろん満たしてくれないことも往々にしてありました。よくあるのは、「日本語に翻訳してください：I love cats」と入れたのに、“Cats are popular pets in Japan”という文が出力されるといった場合です。言語モデルらしい挙動として、続きの文を生成してしまうわけです。そうではなく指示を聞いてもらうために、聞く能力を強化するのが指示チューニングです。

仕組み自体はとても単純です（**スライド 10**）。何かやってほしいタスクの指示と入力文、期待される出力文のセットをたくさん集めてきて言語モデルの訓練を続けると、LLM は指示部分を解釈して望ましい文章を出力するようになります。例えば、「あなたのジョブはこの日本語を英語に訳すことですよ」という指示と、翻訳してほしい日本語の文章を与えると、指示の部分を解釈して言うことを聞いてくれるようになります。このような学習をたくさん行うと、例えば翻訳は学習していても要約は学習したことがないモデルであっても、指示を理解して新たに要約をできるようになります。これが指示チューニングの仕組みです。

しかし、これだけだと、言うことは聞いてくれますが、不適切なことを言うこともあります。LLM は大部分をウェブテキストから学習していると述べましたが、ウェブテキストには人間の社会を反映したような様々なバイアスが入っており、攻撃的な発言や不適切な発言をすることもあります。このような言語モデルを利用する場合、人間の規範や社会規範に合わない出力をするのは望ましくないため、それを抑制する仕組みが取り入れられるようになってきました（**スライド 11**）。どのようにするかというと、まず人間が頑張ってラベルを付けます。例えば、爆弾の作り方を答えたら駄目ですよ、差別的な発言は不適切ですよというラベルをたくさん用意し、それを用いて、人間の評価をまねするモデルを別に作るのです。そして、人間をまねするモデルから信号をもらい、言語モデルを更に訓練します。そのことによって、例えば爆弾の作り方を聞かれたら答えていたような言語モデルが、回答を拒否する能力を身に付けるようになります。指示に従う、社会規範に合わせて答えられるようにするという二つの訓練を追加したのが、現在の ChatGPT のような対話ができる言語モデルです。

このような言語モデルの光の面としては（**スライド 12**）、言語生成能力がとても高くなっていることです。現在は言語モデルが作った文章と人間が作った文章を判別するのが難しいレベルに達しており、高い言語生成能力をいかして様々な人間の業務・作業の補助ができるだろうと期待されています。

例えば、最近のパソコンには幾つかの言語モデルが標準実装されており、クリエーションやリアルタイムキャプション(1)を付けるのを手伝ってくれます(2)。また最近は、検索にも言語モデルが搭載されたことによって検索結果を要約してくれたり、たくさんの情報を見るのを助けてくれたりするようになっています。

コンピュータプログラムが学習データに入っていると述べましたが、プログラムを書くのを助けてくれるサービスも実際に GitHub(3)で使われています。プログラムを書く訓練をしていない一般の方でも、言語モデルと会話をしながらプログラムを書けるような時代になってきています。

---

(1) 動画や音声の内容を即時に文字で表示する機能。
(2) 例えばマイクロソフト（Microsoft）社の「Copilot+PC」など。
(3) ソフトウェア開発のプラットフォームとして広く使われているサービス。

研究の分野でも言語処理の分野でも様々な言語モデルの応用が考えられており、外国語学習を支援するのに使っています（**スライド 13**）。今、中高の先生は業務負担が非常に大きいと言われています。特に外国語学習においては、自分の教えている生徒のレベルに合った教材を作る必要があるため、大変だという面があります。これに対し、教材作りの一部を言語モデルによって自動化して補助することで、ネイティブ話者が書いたような難しい文を、例えば日本人の中学生が理解できるレベルに自動で平易に書き換えて、教材で使いやすいようにして提供する研究をしています。教育においては学生にフィードバックを与えることがとても重要ですが、現在その時間がなかなか取れないため、教材の準備時間を少し減らして、フィードバックにもう少し時間を割けるようにするということへの活用も期待しています。

また、言語モデルは医療文書処理のような専門的な文書の処理にも使えます。医療行為では膨大な記録作業が発生しますが、文章を自動で要約して簡単に患者の情報を見直すことができるよう補助できないかということを今研究している段階です。

一方で、光の部分だけではなく影の部分もあります。皆さんも御存じのとおり、様々な問題が大規模言語モデルにあると指摘されています。**スライド 14** は現在の大規模言語モデルにあるリスクを分類してまとめている樹形図です。よく知られているのは、言語モデルが差別的な発言、有害な発言、攻撃的な発言を生成することや、言語モデルからプライベートな情報が漏れてしまう、あるいは機密情報を言語モデルから取り出せてしまう可能性があることです。さらに、悪用する人が当然出てきます。アメリカ大統領選の時にも言われていますが、偽情報の生成に使われることもあります。現在はとてもリアルな偽情報を作ることができるため、そのようなものに悪用されることが指摘されています。

図の左上にある Human-Chatbot Interaction Harms はとても新しい有害さの懸念です。**スライド 15** にはこの部分を拡大したものを載せました。これは、言語モデルがあまりにも自然な文章を生成するようになってしまい、言語モデルに依存する人が出てくる可能性があるという指摘です。実際にニュースになっているとおり、言語モデルとの会話に傾倒してしまった結果、自殺する人が出ています。依存性を高めるような対話も新たなリスクであるということが知られるようになってきました。

言語モデルができて嬉しいと喜んで研究しているだけではありません。安全性や信頼性をどう保障するのか、リスクをどう減らすのかが、現在非常に活発に研究されている分野の一つです。現時点で調べた範囲では、我々が論文発表をするような国際会議[4]では 176 件、このようなリスクに関する研究が出ています（**スライド 16**）。その論文のタイトルからタグクラウド（**スライド 16** 右側の図）を作ると、例えば、バイアスやアタック、プライバシー、自殺などの単語が出てきます。このような問題をいかに解決していくか研究がなされているところです。

我々もセキュリティや機械学習の先生とチームを組んで、こうした言語モデルのリスクをどう軽減するか、いかに社会に許容されるような言語モデルを開発するか、プロジェクトとして取り組み始めたところです。

（あらせ　ゆき）

---

(4)　Annual Meeting of the Association for Computational Linguistics, Conference on Empirical Methods in Natural Language Processing 等。