より安全な大規模言語モデル（LLM）を目指して

LAWRENCE Carolin　BIFULCO Roberto　GASHTEOVSKI Kiril　HUNG Chia-Chien　BEN RIM Wiem　SHAKER Ammar　小山田 昌史　定政 邦彦　榎本 昌文　竹岡 邦紘

要　旨

大規模言語モデル（LLM）は私たちの世界に革命をもたらしています。LLMは、人間のユーザーが知的システムと対話する方法に根本的な変化をもたらす優れたテキスト能力を備えてはいますが、LLMで作業する際に留意すべき重要な複数の限界も存在します。NECでは、こうした限界に対処するための方法を2つの方向から探りました。第一に、(1) ユースケースのリスクを評価すること、(2) LLMから説明を引き出せるようにプロンプトを与えること、(3) LLMを人間中心のシステム設計基準に収めること、を含めた、既に利用可能な選択肢を検討しました。第二に、(1) 高リスク領域でのLLMの品質をより正確に評価すること、(2) 生成されたLLM出力を入力と結び付けて説明すること、(3) 生成されたLLM出力を外部の信頼できるソースによってファクトチェックすること、などを可能にする現在開発中の技術について報告します。

KeyWords　要約／質問応答／安全性／人間中心／説明可能性

1. はじめに

この目覚ましい技術進化の時代において、大規模言語モデル（Large Language Model、以下、LLM）は革新的な勢力として台頭してきています。多くのタスクで発揮されるその驚異的な性能[1]により、LLMをベースとする技術は現実世界での適用分野を爆発的に増大させています。LLMは、社会とコンピュータシステムとの関係性、テキストによる情報、そして私たち自身の創造性さえ変革しつつあります。さまざまな産業にも革命をもたらそうとしており、ほとんどの専門業種にも適用が可能であり、テキストコンテンツを生成し、仮想アシスタントを向上させ、大規模なデータ解析に基づく洞察をもたらすのに有用です。このようにLLMは、前例のないレベルの利便性と効率をもたらす好機といえます。しかし、NECではLLMが提供する多くの利点を探求したところ、限界があることを認識しつつ、可能なミティゲーション（緩和措置）戦略を考慮しながら慎重に進むべきであることが分かりました。

例えば、LLMにおいて最も議論されている短所は、一般にハルシネーションと呼ばれる、もっともらしく見えながら誤っている情報を生成するという傾向です[2) 3)]。LLMの利点を安全に活用するには、このような限界に注意することが極めて重要となります。一例として、ある法律家が有名なLLMであるChatGPT[4]を、不正確な情報を生成する可能性に気付かず検索エンジンとして使用し、出力を人手で確認することなく書類として法廷に提出したことがありました。

第2章では、まずLLMの主要な長所と能力の実例を示し、続いて最も重要な限界を示します。これに留意したうえで、第3章では、適切なLLMユースケースの全体的特徴について説明し、更には既存の限界が引き起こすリスクを低下させるために今すぐ適用可能なミティゲーション（緩和措置）戦略について検討します。これには、(1) ユースケースのリスクを評価して必要なミティゲーション（緩和措置）戦略の深度を判定すること（第3章1節）、(2) 推論経路についてLLMから自然言語による説明を引き出せるようにプロンプトを与えること（第3章2節）、(3) LLMを人間中心のシステム設計範囲内に収め、人間のユーザーがLLMを安全に利用するために必要な制御と道具を提供し安全な使用を容易にすること（第3章3節）、が含まれます。

第4章では、LLMのユーザビリティを更に向上させるため、現在開発中である、LLM使用の安全性を高められる次の3つの技術に着目します。第4章1節のクオリティチェッカーは、応答の正確さにとどまらない、例えばある

応答にヘイトスピーチが含まれていないかなどの安全性を測定するようなLLM性能評価を、より正確に行えるようにします。第４章２節のLLMエクスプレーナーは、LLMの生成した出力フレーズを、この出力フレーズ生成に使用した入力フレーズとリンクさせることにより、LLM生成出力の説明を可能にします。第４章３節のファクトチェッカーは、外部の信頼できるソースに基づき、生成されたLLM出力の妥当性確認が可能であるかどうかを検証できるようにします。第５章では、エグゼクティブサマリーでまとめます。

2. 大規模言語モデル（LLM）

LLMはテキスト処理のためのニューラルネットワークモデルの一種であり、一般にはTransformerアーキテクチャ[5]を実装したニューラルネットワークをベースとしています。特定のタスクを対象とするラベル付きデータセットで訓練された過去の言語モデルと異なり、LLMは膨大な量のデータを用いた教師なし学習によって訓練します。この訓練の目標は、与えられた入力プロンプトに対し次の単語を予測することです。訓練目標の単純さとラベルなしデータからの学習能力により、LLMは大量のデータを取り込むことでスケールを変えます。この訓練法を拡張することにより、多くの未見のタスクを解決し創発的振る舞いを獲得するというモデルの機能を高めることが、十分に可能だと証明されています[6) 7)]。例えば、LLMは質問応答[8)]、物語生成[9)]、情報抽出[10)]、テキスト要約[11)]などさまざまなタスクを実行できます。更に驚くべきことに、これらの創発的能力には独創的な言語生成や推論、問題解決、そしてドメイン適応[1)]が含まれます。幅広いタスクについて創発的能力を備えていることから、研究コミュニティではしばしば、大規模に事前訓練されたLLMを基盤モデル（FM: Foundation Model）と呼んでいます[12)]。この定義は同一モデルで幅広いタスクを実行できるという能力を反映したものであり、またプロンプトとしてテキストにとどまらず、例えば画像のような他のモダリティも扱えるという、近年の発展にまで基盤モデルのコンセプトを一般化したものです。

2.1 長所

LLMは多くのタスクで素晴らしい可能性を示し、より対象領域を絞ったモデルをしばしば凌駕します。しかし、与えられた特定のタスクを解決するLLMの能力のみを語るだけでは、LLMの最も重要な長所を本当に理解することはできません。実際、LLMの最も重要な点は、適用する際の方法にあります。他の機械学習モデルと異なり、LLMはそのままで目的とするタスクに使用でき、どのような改良も追加の訓練もドメイン固有のデータセットも必要ありません。これにより、自らのシステムに機械学習を統合する際に技術者が直面する技術的障害のほとんどを取り除くことができ、機械学習の適用分野の開発や技術革新をこれまでにないペースで進めることが可能になります。この意味で、LLMの最も興味深い創発能力はテキスト内学習及び指示実行の能力であるといえるでしょう。すなわち、ユーザーはLLMに自然言語による具体的な指示をプロンプトに与えて、LLMの振る舞いをプログラムすることができるのです。このため、機械学習についての専門技術がなくてもLLMを利用でき、更に技術職でないユーザーでも興味ある適用分野でLLMを使用できるようになります。例えば、LLMを高品質なテキスト要約を提供する専用のシステムとして働かせたいなら、「下記のテキストを要約しなさい」といったプロンプトで十分です。その結果、このLLMの能力によって、望みの結果を得るためにLLMにどのようにプロンプトを与えればよいかを探索する「プロンプトエンジニア」と呼ばれる新しい職種が生まれています。

2.1.1 流暢なテキスト

LLMの可能性を支えているのは、さまざまなスタイルや文脈で流暢なテキストを生成できる能力です。LLMは、口語体の散文を、歌詞や詩になるほどの詩的な表現へ、そして法的書類のような特定分野固有の改まった文章へ、簡単に変換できます。複数言語のテキストをシームレスに取り扱う能力と組み合わせることで、どのようなシステムの入出力にも対応する汎用テキストインタフェースが可能になります。例えば、スペイン語で書かれたテキストについて英語で指示を書き、LLMにドイツ語で回答を要求するようなこともできます。同様にLLMに対し、流暢なテキストを提示するのにとどまらず、例えば表やマークダウン記法あるいはHTMLなどでフォーマットされた入出力について、構造フォーマットを統合するように指示することも可能です。

2.1.2 Few-shot能力

指示に従う能力のことを「ゼロショット学習」と呼ぶことがあります。タスクは記述するものの、タスクの例をまったく提供しないからです。あるタスクをLLMに教えるのにもっと指示が必要な場合、いくつかの例をプロンプトとして提供することにより、LLMは目標とするタスクを推論することが可能になります。

例えば、提供したレコードから第二の数字セット“12465”を推論するようLLMに指示するには、次のプロンプトで十分です。

“ record: 235-32446-abc-d
code: 32446
record: 631-12465-lkj-e
code:
”

この能力は一般に「Few（少数・少量）shot学習」と呼ばれており、タスクを言葉で記述すると曖昧になる、あるいは他の方法がユーザーにとって難しい場合、特に便利です。

2.1.3 プログラムコード

LLMは自然言語に加えてプログラミング言語を扱うこともできます。これによって、実用的なプログラムの生成という、興味深い別の可能性が得られます。実際、LLMはプログラムコードに関連する次のような多くの可能性を示しています。

（i）ソースコードを理解・説明できる
（ii）コードスニペットのバグを修正できる
（iii）プログラミング言語とソフトウェアライブラリとの間の変換ができる。例えば、Pythonの関数からC++の関数へ変換する、あるいはNumPyを使用するスクリプトを、同等の機能を持つPyTorch関数を使用するように修正する
（iv）指示を与えられればゼロからプログラムを生成できる
（v）与えられた指示に従って補完や仕上げを行い既存のプログラムを拡張できる

2.2 限界

LLMは大きな潜在能力を備えていますが、同時に顕著な限界も存在します。1つ目は、訓練には大きな費用がかかることから、再訓練する頻度が低くなってしまうことです。このため、LLMは最新の知識にアップデートすることができません。2つ目は、プロンプトの入出力のサイズに通常は上限があることです。LLMの入力はまずトークン化されてからLLMに与えられます。トークンは、単語の一部分のようなものと考えることができます。例えば、あるモデルに対する入出力の合計サイズの上限が4kトークン（約3k～3.5k単語）である場合、処理できる入力の種類も制限されてしまいます。3つ目は、LLMが不正確な出力を生成することがあり、この現象は「ハルシネーション」と呼ばれています[2) 3)]。この場合、LLMの生成した回答が一見信頼できそうに見えても、内容は曖昧であったり、誤っていたり、完全なでっち上げでさえあったりします。

2.2.1 ハルシネーション

プロンプトが与えられると、LLMは常に流暢で自信ありげな応答を生成しますが、生成されたテキストが必ずしも正しいわけではありません。したがって、LLMの深刻な脆弱性は、流暢ではあるが不正確なテキストを生成する傾向を持っていることであり、テキストは一見するともっともらしいが実際には不正確なためにハルシネーションと呼ばれます。

例えば、「フランスは何回サッカーワールドカップで優勝しましたか?」という質問を与えると、LLMは「フランスはワールドカップで1回、1998年に優勝しています」と自信ありげに回答することがあります。しかし実際には、フランスはワールドカップで2回優勝しており、それは1998年と2018年です。これは、あるイベントが除外されるというハルシネーションの一形態であり、そのために流暢で自信ありげだが間違った回答が返ってきたのです。

同様に、「私はフランスが3回優勝したという事実を知っていますよ」と会話を続けたとしましょう。LLMは「謝罪します。あなたが正しい。フランスは1998年、2018年、1958年の3回優勝しています」と答えることがあります。これもまたハルシネーションであり、今回は余計な年をでっち上げて追加しています。そのうえこれは、ユーザーを喜ばせたい、その望みに応えたいというLLM共通の振る舞いを示したものでもあります。

ハルシネーションが特に危険なのは、事実と誤った情報が混じり合った複雑な回答が作成された場合です。この結果として、ユーザーは出力全体を信頼してしまい「権威

に訴える論証」という誤りを犯すことになってしまいます。例えば、医学的アドバイスに関するプロンプトを与えたケースで、回答の末尾に医師に相談すべきというヒントが付加されることがあります。これは有用なヒントではありますが、一般的なことであり、事実か事実でないかという議論との関連が曖昧なため、しばしば深読みによる誤解が生まれます。

2.2.2 複雑な推論の欠如

LLMは人が書いたようなテキストを生成するのには優れていますが、しばしば常識的な理解に欠けることがあります。LLMは訓練に使われたデータが持っている統計的パターンに依存しています。言い換えれば、LLMは与えられたある入力に対し、次のトークン（＝単語）を予測するよう訓練されているのです。これにより、状況によっては事実の間違いや論理的でない応答を返すことがあります。このため、LLMは「stochastic parrots（確率的オウム）」[13]とも呼ばれます。世界についての物理的な理解が欠けているためにLLMが失敗しやすい複雑な推論を伴うタスクとして、ステップが複数あるテーマや算術的、社会的、一時的、あるいはマルチモーダルな推論などが挙げられます[14]。

2.2.3 隠されたバイアス

LLMはしばしば訓練データ内に存在するバイアスを継承することがあり、社会の偏見やステレオタイプの永続化、更には増大につながる可能性があります。こうしたバイアスは、LLMのテキスト生成や意思決定のやり方に影響することもあります。例えば、英語データで訓練されていることがほとんどであるため、多くのLLMが英語圏諸国の文化に適応した出力を行う傾向にあります。同様に、あるLLMが例えばソーシャルメディアのデータなどで訓練されている場合、訓練データ内に存在し得るあらゆる種類の差別的な見方を提示してしまうことがあります。LLM内のバイアスに対応することは重要な課題であり[15]、意図しない結果を減らすには訓練データの人手による注意深い選別や継続的な監視などが必要です。

2.2.4 ブラックパンドラボックス

現在も調査されている問題点に、LLMが持つかもしれない隠れた有害な能力があります。例えば、LLMの訓練に使われたドキュメントがどの程度安全であるかを完全に把握することはできません。これにより、訓練されたLLMは、黒いパンドラの箱のように見えてしまいます。LLMはプロンプトを与えられた際にどのような有害情報を知っているかを明らかにすることをしばしば拒否します。しかし、敵対的プロンプティングを行うことで、このボックスを開き海賊版メディアや自傷的なコンテンツのダウンロードといった有害情報を明らかにできることが確認されています。

3. より安全な利用

現行LLMの長所だけでなく短所をも考慮すると、次のような2つの大きな疑問が生じます。

（1）望ましい適用領域は何か
（2）現行LLMの利用の安全性を高めるには何をすればよいか

次からは、疑問（2）に対処する3つの選択肢に注目します。その選択肢とは、第3章1節では、LLMのユースケースのリスクを評価すること、第3章2節では、LLMに対し、実施した推論について自然言語で説明するよう求めること、第3章3節では、LLMを人間中心のシステムに組み込むこと、というものです。

疑問（1）の「望ましい適用領域は何か」については、追加データの取り込みや適用分野に特化した訓練をしなくても、LLMが利用できることに留意しなくてはなりません。広い適用分野のなかから、LLMが容易に成功できるものをどのようにして選んだらよいでしょうか。この点では、LLMは偉大で創造的な執筆者であるものの、誤った情報を作り出す可能性があることを考慮する必要があります。つまり、正確さが必ずしも問題にはならない適用分野（例えば小説を書く場合）、または人間が積極的に関与することでミティゲーション（緩和措置）戦略が可能な適用分野であれば、LLMはすぐに利用できるということです。

例えば、法律テキストに役立つLLMの例などは正確さが明らかに必要なケースであり、同時に前述のような懸念を無視できる例といえます。実際、多くの適用分野では、人間の専門家でも誤りを犯すことを認識して中間チェックの仕組みが作られています。これらの適用分野ではLLMは初期テキスト生成を行い、適用分野本来のワークフロー内では人間の専門家とシームレスにペアを組むことができ

ます。法律テキストを例にすると、テキストの正確さを2人の弁護士に相互チェックさせていたような場合、今であれば、ほとんどの「重労働」をこなすLLMを追加することで、うまくすれば人間の作業を単純ミスのチェックだけに減らせると思われます。第3章の以降の節では、LLMにふさわしい適用領域をより正しく判断するための知的ツールをいくつか取り上げます。

3.1 リスク分類

LLMを安全に使用するためのミティゲーション（緩和措置）戦略はLLMを展開するユースケースに大きく依存します。例えば、LLMを本の推薦に使用する場合、LLMが実際には存在しない本をハルシネーションで取り上げたとしても重大な弊害はないので、LLMの低リスクな使用ということになります。これに対し、LLMを患者のための診断書の生成に使用することには高いリスクが伴います。もしその診断書にハルシネーションが含まれており、医師がその診断書に基づいた決定を行った場合、患者には間違っているか、危険でさえある処置が行われてしまう可能性があるからです。したがって、まずユースケースのリスクを評価し、その結果に基づいて適切なミティゲーション（緩和措置）戦略に取ることが肝要となります。

ユースケースごとのリスクレベルの判定には、例えば欧州連合（EU）が制定しようとしているAI規制法（以下、EU AI Act）[16]の定めるリスク定義などを利用できます。EU AI Actでは、AIの使用を4つのリスクカテゴリで分類しようとしています。この法律はまだ完成していませんが、法律や医学などの特定のドメインではLLMに高いリスクがあるとみなされることが予想されます。各リスクカテゴリはそれぞれに用途への影響やAIシステムの展開前に実施すべきチェックを定めており、その概要を**図1**に示します。4つのカテゴリは次の通りです。

（1）最小リスク：AIの利用についての情報をユーザーに提供しなければならず、使用のオプトアウトを選択できる
（2）限定リスク：透明性が要求される
（3）高リスク：EU内でAIシステムを展開する前に適合性評価を実施する必要がある
（4）許容できないリスク：AIのユースケースでの使用は許可されない

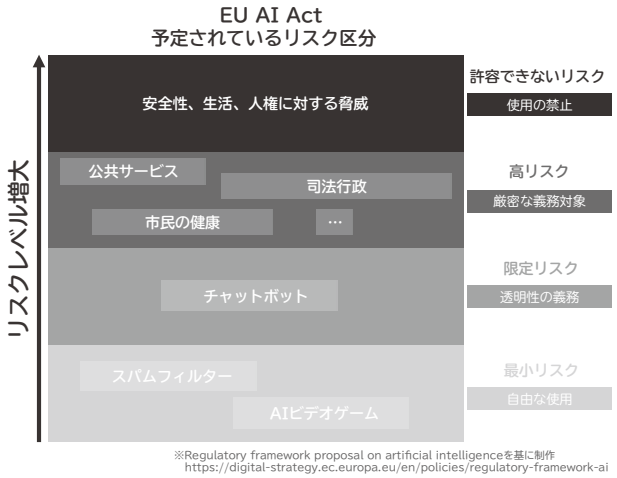

図1　EU AI Actにおける、リスクレベルに従って区分けしたAIの適用分野

3.2 自然言語による説明

ミティゲーション（緩和措置）戦略の1つとして、単純にLLMに対して応答とともに推論経路を生成するよう要求することができます。LLMへのプロンプトのなかに、推論の背景についての説明を自然言語で提供するよう明確に尋ねる要求を盛り込むことができます。例えば、「私はバスケットにリンゴを3個持っています。スーパーマーケットへ行ってリンゴを6個買いバスケットに入れました。帰り道で2人の友人に会ったのでそれぞれに1個ずつリンゴをあげました。家に帰ったとき、バスケットには何個のリンゴが残っているでしょうか」というテキストベースの計算問題を解くようにLLMに求めるというシナリオがあるとします。

このプロンプトに基づき、正しい解答は「7個」ですが「あなたのバスケットには4個のリンゴが残っています」とLLMが誤った解答をすることがあります。人間にはなぜAIがこのような解答をしたか理解できず、確認するためにできることといえば、自分自身で問題を解く他ありません（これでは、まずLLMに聞くことの利点がなくなるでしょう）。しかし、入力に単純な変更を加えるだけで、LLMに対して解答に加えて推論も生成するよう依頼できます。例えば、元の入力に「段階を踏んで考えてください」という文章を入れれば推論経路を得られます。この例の場合であれば、次のような応答を受け取ることができます。「バスケットのリンゴはまず3個ありました。スーパーマーケットであと6個のリンゴを買ってバスケットに入れると、合計は

3＋6＝9個になります。帰り道で2人の友人に1人1個ずつリンゴをあげると、結果として9－2＝7個のリンゴがバスケットに残ります」

推論経路を得ることができたのに加え、今度こそLLMが正しい解答にたどり着くという利益も得られました。LLMに最新の推論のステップを繰り返しプロンプトに与えることで、正確な回答を作り出せるようLLMを暗黙的に導くことができます。

このアプローチの潜在的な難点は、今のところプロンプト作成の技術が非常に不安定なことです。これは、特定の句を追加すれば正確な回答が得られると保証できるわけではなく、フレーズをほんの少し変える（例えば句の最後に単純に！マークを追加する）だけで再び誤った結果になってしまう可能性があることを意味します。

3.3 人間中心アプローチ

現行の弱点そして一般的な統計的性質からすると、LLMが100%正しく機能することを保証するのは不可能です。したがって、LLM利用の安全性を高めるのに何ができるかという疑問が浮かび上がってきます。ここでは、LLMを組み込んだアプリケーションを、人間を中心とする方法で設計する[17]ことにより、人間が制御できる量が増え、より安全な利用につながるということの概要を説明します。

コンピュータアプリケーションが提供する自動化の量は、通常、低から高までの1つの軸で示すことができます。しかしこの見方は、利用する際に人間が制御する量を表す軸を追加することで拡張できます[17]。この枠組みをLLMに適用することで[18]、次のようなシナリオが考えられます

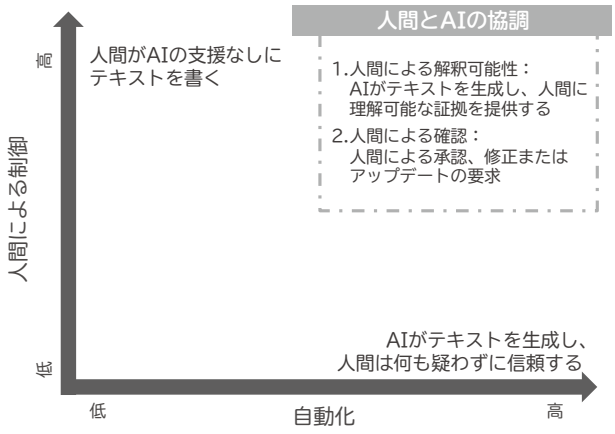

図2　人間による制御でのLLM適用分野の安全性

（**図2**）。LLMが存在しないシナリオでは、人間がテキスト記述のすべてを制御します。人間ユーザーが何も疑わずに信頼するテキストをLLMが生成するシナリオでは、制御が不在になり、ユーザーは現在のLLMの限界及びその結果としての危険に完全にさらされることになります。こうした状況は、人間に正しいツールを提供する第三及び最後のシナリオによって緩和でき、人間による制御が取り戻せるようになります。

NECは人間による解釈と検証を可能にする手段を提供することで、LLMが生成するテキストを人間ユーザーが制御できるようにします。人間による解釈のためには、LLMの生成した出力を人間に理解できるようにする証拠を供給します。このシナリオについては、LLMに対し自然言語による説明を提供するよう求めるという選択肢のことを既に述べています。また、回答を生成する前に関連したテキストの一節を検索するという、検索をベースとする新しいタイプのLLMを検討することもできます[19) 20)]。ただしこれまでのところ性能が基準に達しておらず、まだ広く利用されるまでにはなっていません。

人間による検証を可能にするために、NECではLLMの生成したコンテンツを人間のユーザーが検証できるようにするツールを提供しています。第4章では、その3つの技術を紹介します。

4. より安全な技術

NECは、LLM技術を使う人が、LLMの出力を検証して、その結果LLMをより安全に使用できるような能力を提供できます。次からは、この目標への助けとなる3つの技術を紹介します（**図3**）。

第4章1節では、クオリティチェッカーを紹介します。LLMの正確さが十分かどうかを確認するために、LLM導入前に実行します。正確さという言葉はさまざまなやり方で判断でき、一連の出力がどの程度安全かまたは事実であるかを測定する評価基準を利用することもその一例です。これにより、最低限の品質が保証され、さまざまなLLMを比較してユースケースに最適なものを選ぶことが可能になります。

第4章2節では、LLMのテキスト生成後に実行することができるLLMエクスプレーナーについて紹介します。これは、生成したテキスト内のフレーズとLLMに与えられ

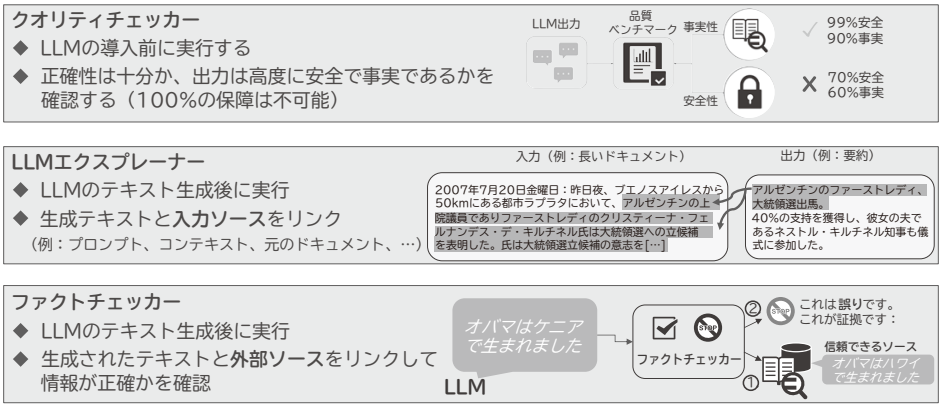

図３　LLM使用をより安全にするため現在開発中の３技術の概要

た入力ソースとをリンクします。例えば、テキスト要約のケースでこれを使用すると、LLMがオリジナルドキュメントのどの一節から要約の文またはフレーズを生成したのかが理解できます。したがって、この技術によると、例えば医師が診断書の正確さを効率的に確認し、自分で診断書を書く時間から解放され他の仕事をする、といったことが可能になります。

第４章３節では、ファクトチェッカーを使用することで、LLMの生成したテキストを外部の（信頼できる）ソースで検証できることを紹介します。これによってユーザーに対してハルシネーションの可能性を警告できます。別のシナリオとして、この技術はフェイクニュースの識別にも使用できる可能性があります。

4.1 クオリティチェッカー

LLMはエラーやバイアスに影響を受けないわけではないため、特にリスクの高いドメイン（例えば図１に示したEU AI Actの通り、法律や医学関係）で使用する際にはこの欠点によって有害な結果が生じることがあります。したがって、LLMを導入する前には、具体的なユースケースに応じたLLMの綿密な評価を行うことが極めて重要です。LLMの評価は標準的に次の２つの側面で行います。

（i）判断に適切なデータセットの選択

（ii）評価手法の確立

前者は評価に適切なベンチマークを特定することであり、後者は自動化の手法に対する評価と人間中心の手法に対する評価の両方を判断するための基準を定義することです[21]。

高リスクなドメインの環境において、LLMの使用に伴う複雑さとその潜在的な影響があることは、より包括的で決定的な評価プロセスが必要であることを示しています。LLMを評価する際には、具体的な課題があります[22]。例えば法律などの分野では、内容の適切さを保つために定期的なアップデートが求められます[23]。医療の分野では安全を最優先にして判断が行われなくてはならず、発生する可能性が高いハルシネーションによって有害な結果が生まれる恐れが潜在的にあるため、現在の適用範囲は厳しく制限されます[24]。

このことにより、LLMの性能を評価する際には、事実の正確さと安全性に関する懸念の両方に取り組むことの重要性が際立ちます。NECでは以前、ドメインに適合するインストラクションチューニングを施したLLMが、高リスクドメインのタスクにおいてどのように働くか（つまり質問応答と要約を行うか）について、HUNG Chia-Chienらが法律と医学の分野で調査しました。そのために、「事実性」[25) 26)]と「安全性」[27) 28)]を測定できる最先端の評価基準に関して、さまざまなLLMの性能を各種の高リスクテストのセットで計測できるクオリティチェッカーを作成しました。調査の結果として、高リスクドメインのタスクとLLMの適合性には大きなギャップがあることが明らかになり、その時点では人間中心のアプリケーションに注意深く組み込まない限り、LLMの使用はまだ実用的ではないということが示唆されました。

全体的に、現在実装されている評価基準を備えたクオリティチェッカーは、そのLLMがどのようにうまく機能するか、またあるユースケースに最良のLLMはどれかを示す

ことはできます。しかし、(1) 特定分野のアプリケーションに適合した評価基準を定義し[29]、(2) モデル出力の正確さを最もうまく評価し、安全性に関する懸念にも対応する方法について、分野の専門家と協力して調査する[30]ためには、更なる努力が求められます。したがって、現時点においてクオリティチェッカーは、人間中心のアプリケーション設計など、他のミティゲーション（緩和措置）戦略と組み合わせる必要があります。

4.2 LLMエクスプレーナー

ハルシネーションに陥りやすいために、LLMを医療などの高リスクドメインで直接利用することが困難になっています。例えば、LLMがある医療指示の要約として「患者は薬品Xの50mg錠剤を、1日3回摂取すること」という文を作成したとします。しかし、元のドキュメントを調べると、患者は薬品Xの5mg錠剤を、1日1回摂取するよう推奨されているので、これはハルシネーションであると分かります。この例では要約のハルシネーションは患者に本来の用量の30倍の薬品を摂取するよう推奨しています。もし患者が何の疑問も持たずにLLMを信じ、生成された要約の指示に従った場合、患者にとっては深刻な結果（例えば、致死性のオーバードーズ）につながることがあります。

NECのLLMエクスプレーナーはこうした問題を避けられるよう、生成されたLLMテキストとオリジナルの入力間にリンクを作成できます。ここではLLMエクスプレーナーがリンクすべき情報は、既にLLMへの入力クエリの一部になっていると仮定します。例えば、このエクスプレーナーはLLMが要約した文からオリジナルのソースへのマッピングが可能です（図3中段）。このリンクにより、ユーザーはLLMが生成した情報の正確さを効率的に確認できるようになります。同様に、どの情報がオリジナルの入力には存在して要約には存在しないかを強調することが可能です。

このエクスプレーナーはまた、オリジナルのテキストを利用する他のタスクにも使用できます。例えば質問応答では、質問及びその質問に関連があって回答の基礎とするべき入力テキストを、ユーザーが提供します。このエクスプレーナーでは、生成された回答から、入力テキスト内でその回答を参照している箇所への逆マッピングが行えます。

全体として、エクスプレーナーは、どの入力フレーズがどの出力フレーズの生成をもたらしたのかをユーザーが理解できるようにするツールです。これにより、ユーザーは出力の正しさを確認でき、高リスクドメインでLLMを使用できるようになります。この場合ユーザーは、コンテンツ生成をLLMにアウトソーシングすることで時間を節約できる可能性がありますが、安全上の理由から出力の確認が必要です。対照的に、ユースケースによっては、関連情報が入力クエリとともに提供されないこともあります。

このシナリオでは、LLMは代わりに明白には表に出ない自らの内部知識にアクセスするため、LLMエクスプレーナーを直接適用することができません。こうした場合は、信頼できるソースのセットとの比較によってLLMの生成情報を検証できるNECのファクトチェッカーに頼ることができます。

4.3 ファクトチェッカー

自動ファクトチェックは、間違っているかミスリーディングな情報を迅速かつ正確に識別できるようにします。さまざまな分野やアプリケーションでのLLM利用が増えたことで、研究者もユーザーもモデルが生成する間違いやハルシネーションを検出できる重要な手段として、ファクトチェックが登場しました。更に、偽情報の拡散がこれまでにないほど簡単で結果も重大になっている時に、伝統的なファクトチェックはソーシャルメディアへの投稿の分析やフェイクニュースの嘘を暴くためにも重要です。

自動ファクトチェックのパイプラインは、次のように説明できます。与えられたテキストをフレーズに分解し、ファクトチェッカーはまずどのフレーズにファクトチェックが必要かを識別します。例えば「Dear ladies and gentlemen（皆様）」にはファクトチェックは不要です。第二に、あるフレーズにチェックが必要となった場合、それは「主張」となり、システムはこの主張に関係した関連性のある参照ドキュメントを検索します。この結果に基づき、第三そして最終のステップで、ファクトチェッカーはその主張の真偽を判定します。

標準的なファクトチェッカーは、ある主張をなぜ真または偽に分類したか説明できません。これは、分類を信頼できるようにするには依然として人間が確認を行わなければならない、その結果ファクトチェッカーの恩恵は大いに低下してしまう、ということを意味します。そうではなく、ファクトチェッカーが推論経路も出力できれば、システムの有用性を高めることが可能です。例えば**図4**のように、「NECは1899年に東京大学出身者により設立されまし

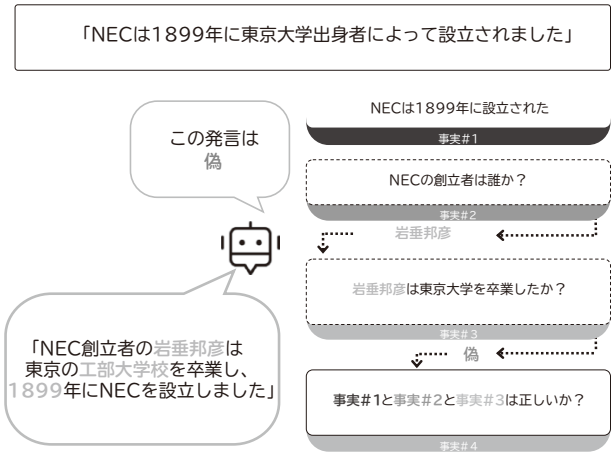

図4　ファクトチェッカーの例

た」という主張は、岩垂邦彦は東京大学ではなく工部大学校を卒業しているため誤りであると識別しています。推論経路を提供することで人間はファクトチェッカーの決定の裏にある理由を理解できます。

推論経路の供給は、重要な第一歩です。これに続いてNECでは、次のような2つの側面からファクトチェッカーの更なる改良を計画しています。第一の側面は、現行システムの推論経路[31]は主張の真偽判定に使用した証拠ドキュメントを提供していますが、その証拠が主張を拒否あるいは支持している理由や方法を説明していないということです。このような説明を追加すれば、人間のユーザーの確認プロセスをスピードアップでき、システムの有用性は更に強化されます。第二は、既存の多くのファクトチェック用ベンチマークは、分野と有用性がしばしば限定されてしまうということです。例えば、多くのベンチマークでは正しい証拠ドキュメントが既に用意されていることを前提としていますが、実際にはまずそれを見つける必要があります。ファクトチェッカーを必要とする実際のユーザーと関わることは、より現実的な調整を進めるうえで助けとなっています。

5. エグゼクティブサマリー

LLMは、世界に革命を起こしています。LLMは、(1) 流暢なテキストを書くこと、(2) わずかのデモンストレーションによって新しいテキストベースのタスクを学習できること、(3) プログラムコードのスニペットを書くこと、を含む優れた能力を備えています。こうした強みに加え、次のような現在の限界についても検討しました。

(1) LLMはハルシネーションに陥ることがあり、これは、実際には誤っているのに流暢で自信ありげなテキストを生産することを意味します。
(2) LLMは世界認識と常識に欠けており、したがって複雑な推論はできません。
(3) LLMは既存のテキストデータにより訓練されているので、データがバイアスを含んでいたりそのバイアスを増幅したりする可能性があります。

この初期評価に基づき、どうすれば現行のLLMの安全な利用を拡大できるかという問題に目を向けました。これについては、次の3つのアプローチを確認しました。

(1) LLMを利用するユースケースのリスクレベルを分類できれば、信頼でき安全な応答をLLMから確実に得るためにどの程度の注意が必要かを知ることができます。
(2) LLMから回答を得るためのプロンプトの与え方を修正できれば、例えば「段階を踏んで考えてください」と加えることで、LLMは推論経路を自然言語で生成できるようになるかもしれません。
(3) LLMアプリケーションの設計で注意できれば、人間が最終的な制御を保持するようにできます。例えばNECのLLMエクスプレーナーを使用して出力を効率よく検証することで、医師は診断書をより早く、しかも安全に書くための支援をLLMから受けられます。

最後に、LLMをより安全に使用するための助けとなる、現在開発中の3つの技術について説明しました。

(1) クオリティチェッカーを使用することで、LLMがあるユースケースについてどのように安全にまたは事実に基づいて動作するかを測定できます。これにより、選択肢のなかから最良のLLMを選び出し、それが十分うまく動作するか測定できます。NECの最初のプロトタイプは、医学と法律の分野で英語によるテストを行いました。将来的には、より具体的な測定事項、例えば、ある応答が医学分野において安全か測定する方法などを考案し、またこれらの測定事項を他の言語でも動作するよう拡張することで、このクオリティチェッカーを

より正確にすることを計画しています。

（2）NECのLLMエクスプレーナーは、LLMの生成した出力をLLMに与えられた入力プロンプトまで遡ることのできる逆トレースが行えます。これにより、例えば要約をより安全に行うことなどにLLMを使用できるようになります。このエクスプレーナーがなければ、要約に間違った情報が紛れ込むかもしれません。またこのエクスプレーナーがあれば、ユーザーは要約のなかの情報は正しいか、重要事項で欠けているものはないかを素早く、効率的に確認できます。これにより、例えば医療報告書作成などの高リスクなドメインでLLMを使用できるようになります。

（3）ファクトチェッカーはテキスト間の矛盾の自動検出に使用できます。入力されるテキストはLLMが生成したものか、あるいは人間が書いたものです。NECのファクトチェッカーは、提供されたテキストを信頼できるテキストソースで構成された参照データベースと照合し、両者の間で矛盾を検出した時には警告を発します。これにより、例えばフェイクニュースの検出などが行えます。

全体として、LLMは大いに有望であり、理解可能で助けになるAIシステムが人間の能力を拡張してくれるという未来をうかがわせます。これによって、以前には考えられなかったような新たな可能性を拓く人間とコンピュータ間の協力が生まれる可能性があります。ただし、生産性向上のためにLLMの潜在力を用いるときには、注意を払わなければなりません。改革をもたらす他の技術と同じように、LLMには固有の限界があり、責任ある利用が必要です。その限界に注意を払いながらLLMの優れた能力を受け入れることで、私たちはともにより明るくより公正な未来への進路を取ることができるでしょう。

参考文献

1) Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean and William Fedus: Emergent Abilities of Large Language Models, Transactions on Machine Learning Research (TMLR), 2022
https://arxiv.org/abs/2206.07682
2) Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro: Factuality Enhanced Language Models for Open-Ended Text Generation, Neural Information Processing Systems, 2022
https://arxiv.org/abs/2206.04624
3) Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto and Pascale Fung: Survey of Hallucination in Natural Language Generation, ACM Computing Surveys, Volume 55, Issue 12, pp.1-38, 2023
https://doi.org/10.1145/3571730
4) OpenAI: Introducing ChatGPT, 2023.6
https://openai.com/blog/chatgpt
5) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz Kaiser and Illia Polosukhin: Attention Is All You Need, Neural Information Processing Systems, 2017
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
6) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al.: Language Models are Few-Shot Learners, Neural Information Processing Systems, 2020
https://arxiv.org/abs/2005.14165
7) Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo and Yusuke Iwasawa: Large Language Models are Zero-Shot Reasoners, NeurIPS2022, 2022
https://arxiv.org/abs/2205.11916
8) Md Adnan Arefeen, Biplob Debnath and Srimat Chakradhar: LeanContext: Cost-Efficient Domain-Specific Question Answering Using LLMs, 2023
https://arxiv.org/abs/2309.00841
9) Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Mohammad Shoeybi and Bryan Catanzaro: Factuality Enhanced Language Models for Open-Ended Text Generation, The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022
https://arxiv.org/abs/2206.04624
10) Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Jiaqi Mu, Hao Zhang and Nan Hua: LMDX: Language Model-based Document Information Extraction and Localization,2023
https://arxiv.org/abs/2309.10952

11) Haopeng Zhang, Xiao Liu and Jiawei Zhang: SummIt: Iterative Text Summarization via ChatGPT, 2023 https://arxiv.org/abs/2305.14835
12) Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill et al.: On the Opportunities and Risks of Foundation Models, 2021 https://arxiv.org/abs/2108.07258
13) Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, The 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ' 21, pp.610-623, 2021 https://dl.acm.org/doi/10.1145/3442188.3445922
14) Wenting Zhao, Mor Geva, Bill Yuchen Lin, Michihiro Yasunaga, Aman Madaan and Tao Yu: Acl 2023 tutorial: Complex Reasoning in Natural Language. ACL 2023, pp.11-20, 2023 https://aclanthology.org/2023.acl-tutorials.2/
15) Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla and Oskar Van Der Wal: You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings, BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, pp.26-41, 2022 https://aclanthology.org/2022.bigscience-1.3/
16) European Commission: Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEG-ISLATIVE ACTS, 2021 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
17) Ben Shneiderman: Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, International Journal of Human–Computer Interaction, Volume 36 Issue 6, pp.495-504, 2020 https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118
18) Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner and Carolin Lawrence: Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains, EMNLP 2023 Workshop on Benchmarking Generalisation in NLP (GenBench), 2023 https://arxiv.org/abs/2311.14966
19) Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat and Mingwei Chang: Retrieval Augmented Language Model Pre-Training, The 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp.3929-3938, 2020 https://proceedings.mlr.press/v119/guu20a.html
20) Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen and Laurent Sifre: Improving Language Models by Retrieving from Trillions of Tokens, The 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp.2206-2240, 2022 https://proceedings.mlr.press/v162/borgeaud22a.html
21) Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang et al.: A Survey on Evaluation of Large Language Models, 2023 https://arxiv.org/abs/2307.03109
22) Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu and Robert McHardy: Challenges and Applications of Large Language Models,2023 https://arxiv.org/abs/2307.10169
23) Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky and Daniel Ho: Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, Neural Information Processing Systems 35 (NeurIPS 2022), pp29217-29234, 2022 https://proceedings.neurips.cc/paper_files/paper/2022/hash/bc218a0c656e49d4b086975a9c785f47-Abstract-Datasets_and_Benchmarks.html
24) Sandeep Reddy: Evaluating large language models for use in healthcare: A framework for translational value assessment, Informatics in Medicine Unlocked, Volume 41, Article.101304, 2023 https://www.sciencedirect.com/science/article/pii/S2352914823001508
25) Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong: QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization, The 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.2587-2601, 2022 https://aclanthology.org/2022.naacl-main.187/
26) Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji and Jiawei Han: Towards a Unified Multi-Dimensional Evaluator for Text Generation, The 2022 Conference on Empirical Methods in Natural Language Processing, pp.2023-2038, 2022 https://aclanthology.org/2022.emnlp-main.131/

27) Laura Hanu and Unitary team: Detoxify
https://github.com/unitaryai/detoxify
28) Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau and Verena Rieser: SafetyKit: First Aid for Measuring Safety in Open-Domain Conversational Systems, The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.4113-4133, 2022
https://aclanthology.org/2022.acl-long.284/
29) Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi et al.: Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity, 2023
https://arxiv.org/abs/2310.07521
30) Xiang' Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K.Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl D.D. Willis, Chien-Sheng Wu et al.: Next Steps for Human-Centered Generative AI: A Technical Perspective, 2023
https://arxiv.org/abs/2306.15774
31) Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan and Preslav Nakov: Fact-Checking Complex Claims with Program-Guided Reasoning, The 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.6981-7004, 2023
https://aclanthology.org/2023.acl-long.386/

執筆者プロフィール

LAWRENCE Carolin
NEC Laboratories Europe
Manager

BIFULCO Roberto
NEC Laboratories Europe
Manager

GASHTEOVSKI Kiril
NEC Laboratories Europe

HUNG Chia-Chien
NEC Laboratories Europe

BEN RIM Wiem
NEC Laboratories Europe

SHAKER Ammar
NEC Laboratories Europe

小山田 昌史
データサイエンスラボラトリー
主席研究員・研究グループ長

定政 邦彦
データサイエンスラボラトリー
プロフェッショナル

榎本 昌文
データサイエンスラボラトリー

竹岡 邦紘
データサイエンスラボラトリー
特別研究員
主任

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報（日本語）　NEC Technical Journal（英語）

Vol.75 No.2　ビジネスの常識を変える生成AI特集
～社会実装に向けた取り組みと、それを支える生成AI技術～

ビジネスの常識を変える生成AI特集によせて
生成AI技術への取り組み ～基盤から応用、ルール作りまで～

◇ 特集論文

急速に広まる生成AIの市場適用
NECにおける生成AIの取り組みについて
電子カルテと医療文書の作成支援による医師業務効率化
映像×LLMによる報告書作成業務の自動化
映像分析と生成AIによるリアル世界の行動理解
サイバー脅威インテリジェンス生成自動化
生成AIの社内活用を進めるNEC Generative AI Service（NGS）
ソフトウェア・システム開発への生成AIの活用
LLMとMIで革新する素材開発プラットフォーム
LLMと画像分析を活用した被災状況の把握

生成AIの可能性を高める基盤技術
日本語性能に優れたNECの大規模言語モデル（LLM）
NECの生成AIを支える国内企業で最大規模のAIスーパーコンピュータ
より安全な大規模言語モデル（LLM）を目指して
データを秘匿したまま連携できる連合学習技術とLLMへの適用可能性
大規模言語モデル（LLM）によるFew-shotクラスタリングの可能性
オープンドメイン常識推論のための知識拡張型プロンプト学習
AI連携とオーケストレーションのための基盤ビジョンLLM
クエリを考慮した新規手法により関連する企業データを減らしLLM APIの使用コストを最適化

AI技術が社会へ浸透するために
AI標準化・ルールメイクに関する動向とNECの取り組み
人権尊重に向けたNECのAIガバナンスの取り組み
RCModelを用いたAIリスクマネジメントのための人材育成事例

◇ NEC Information
2023年度C&C賞表彰式典開催

Vol.75 No.2
（2024年3月）

特集TOP