堀のある AI から責任ある AI へ

生成 AI のリスクを最小限に抑える

Huma Abidi インテル コーポレーション AI ソフトウェア製品ジェネラル・マネージャー兼シニア・ディレクター
Haihao Shen インテル コーポレーション AI ソフトウェア・アーキテクト

OpenAI の ChatGPT* は、記録的な成長を遂げ、生成 AI の世界に熱狂を引き起こしました。なんと、最初の 1 週間で 100 万人以上のユーザーを魅了しました。Google、Microsoft、Meta（旧称 : Facebook）などの大手テクノロジー企業が大規模言語モデル（LLM）レースに参加する一方で、小規模なスタートアップ企業も前進を続けています。競争上の優位性のために必要な秘密性と安全のために必要な透明性のバランスを取ることがますます大きな問題となっています。一部の LLM とは異なり、OpenAI はトレーニング・セットや GPT-4* アーキテクチャーの詳細を公開していないため、一部から批判を浴びています。

最近流出した Google のメモ（英語）では、Google と OpenAI に LLM テクノロジーにおける競争上の優位性、つまり「堀」が欠けていることへの懸念が強調されています。このメモでは、オープンソースの代替品がいかに小さく、速く、安価で、カスタマイズ可能であるかが強調されていました（この点については、本号の「**独自のカスタム・チャットボットの作成**」で説明します）。

ChatGPT* のような LLM は優れた機能を実証していますが、生成 AI モデルの出現によるマイナス面についての懸念も生じています。大規模モデルと小規模モデル、オープンシステムとクローズドシステムに関する継続的な議論を行うことは重要ですが、パフォーマンスと精度だけが重要ではないことを認識する必要があります。公平性、説明可能性、持続可能性、プライバシーなどの要素も考慮する必要があります。責任ある AI の実践を守ることが、AI の社会的な価値を最終的に決定します。おそらく、AI アプリケーションが倫理的ベスト・プラクティスに準拠することを義務付ける多くの規制が今後導入されるでしょう。今年後半に可決予定の次期 EU 人工知能法(英語)が AI システムを管理する世界初の規制となる見込みです。この法律は、透明性とリスク管理に関するガイドラインを導入することにより、AI に対して人間中心の倫理的なアプローチを促進することを目的としています。

生成 AI モデルはインターネット上で利用可能な膨大な量のデータから学習しますが、これらのデータには社会に存在するバイアス（偏見や先入観など）も含まれており、これらのバイアスを意図せず永続させたり増幅させたりする可能性があります。LLM が誤った情報、フィッシング・メール、ソーシャル・エンジニアリング攻撃などを生成または拡散するために操作されることも考えられます。また、悪意のある攻撃者が意図的に偏った情報や虚偽の情報を使用してモデルをトレーニングすることで、誤解を招くコンテンツが大規模に拡散する可能性もあります。このようなモデルを使用すると、説得力のあるディープフェイクのビデオやオーディオコンテンツを作成できます。例えば、最近 AI が生成したアメリカ国防総省で爆発という偽画像（英語）が急速に広まりました。このようなフェイクニュースは、今後の米国の選挙に対する大きな脅威として浮上しています。この記事で述べられているブルッキングス研究所テクノロジー・イノベーション・センターのシニアフェローである Darrell West 氏の懸念は現実的なものです。

LLM はしばしば「幻覚」を引き起こし、不正確な情報を生成することがあります。これは、モデルが診断や治療の決定に影響を与え、患者に害を及ぼす可能性があるヘルスケアなどの業界では特に深刻な問題となります。AI が幻覚を引き起こすことは既知の現象であるにもかかわらず、LLM を使用し続け、LLM の情報を疑うことなく受け入れる人は後を絶ちません。最近の例では、弁護士が裁判の準備書面として提出した判例の要約が存在しなかったことから、調査に ChatGPT* を使用していたことが判明しました。これらの架空の判例は ChatGPT* により作成されたものでした（英語）。

責任ある AI を求める声はこれまでになく高まっています。コモンウェルス・クラブでのインタビューで、Stuart Russel 教授は ChatGPT* について「ある意味、我々はインフォームド・コンセントをまったく得ることなく、人類に対して大規模な実験を行っていると言える」と述べています。Russell 教授や Elon Musk 氏を含む 1,000 人以上の AI エキスパートからなるグループは、LLM のデプロイを一旦停止するように求めています（英語）。議員、業界のリーダー、研究者たちは、AI の安全なデプロイを確保するため、AI の周りにガードレールを設置して、厳格な規制を構築することの必要性に同意しています。業界のリーダーたちは、AI テクノロジーが人類存続の脅威となる可能性があることを認めています。Center for AI Safety は、「AI による絶滅のリスクを軽減することは、パンデミックや核戦争などのほかの社会規模のリスクと並ぶ世界的な優先事項であるべきである」という声明(英語)を発表しています。この声明には、Geoffrey Hinton（トロント大学名誉教授)、Yoshua Bengio（モントリオール大学教授)、Demis Hassabis(Google DeepMind CEO)、Sam Altman(OpenAI CEO)、Dario Amodei(Anthropic CEO)（敬称略）を始めとする、AI 分野の著名人が署名しました。AI のリスクを軽減するには、責任ある開発、慎重なデータセットのキュレーション、継続的な研究、堅牢な倫理ガイドラインが不可欠です。バイアスに対処し、誤った情報を排除し、ユーザーのプライバシーを保護するには、LLM の透明性と説明責任を確保し、定期的な監査を行うことが重要です。

AI テクノロジーに取り組んでいる企業や個人が、AI の行動規範に従ってソフトウェアが開発およびデプロイされていることを確認する必要があることは明らかです。オープンソースのインテル® Explainable AI Tools（英語）を使用すると、ユーザーは事後モデルの蒸留と可視化を実行して、TensorFlow* モデルと PyTorch* モデルの両方の動作を予測することができます。これらのモデルは、ユーザーが公平性と解釈可能性の問題を検出して防ぐことができるように設計されています。例えば、ユーザーは、インテル® Explainable AI Tools に含まれるオープンソースの Python* モジュール、モデル・カード・ジェネレーター（英語）を使用して、モデルの詳細と、TensorFlow* モデルと PyTorch* モデルの両方のパフォーマンスと公平性のメトリックを表示する定量分析を含む、インタラクティブな HTML レポートを作成できます。これらのモデルカードは、従来のエンドツーエンド・プラットフォームの一部として使用でき、透明性、公平性、説明責任を促進する、表、画像、テキストデータの ML パイプラインをデプロイします。

LLM は通常、大規模な公開データセットでトレーニングされた後、機密性の高いデータ（金融や医療など）で微調整されます。Open Federated Learning（OpenFL）（英語）のようなテクノロジーには機密計算（英語）が組み込まれており、機密データに基づいて LLM を安全に微調整できるため、幻覚やバイアスを減らしながらモデルの一般化可能性を向上させることができます。

AI は、重要な専門知識が不足している、経済的に不利な分野を支援できる可能性を秘めています。現在、LLM の実行には膨大な計算能力が必要であり、通常はクラウドや複数のアクセラレーターを搭載した高価なオンプレミス・サーバーで実行されます。我々は、クラウドに接続できない環境や低コストのエッジ・コンピューティング・デバイスでも高度な AI 技術を利用できるように、LLM の計算の複雑さを軽減し、LLM ベースの推論の効率を高めることに取り組んでいます。