

タイトル：ChatGPT、BARD、その他の大規模言語モデルと規制のある製薬業界との邂逅

著者：Frederick Blumenthal, Martin Heitmann, Stefan Münch, and Brandi Stockton  
(Pharmaceutical Engineering, 2023, Vol 43, No 4, 58-64)

翻訳： 京都大学大学院医学研究科薬剤疫学分野 大学院生 名前 劉 湔 (LIU HAO)

## はじめに

ChatGPT やその他の大規模言語モデルは、世界を変える位置づけにある。また、規制産業全般における機械学習ソリューションの受け入れと普及を変えることができる。しかし、それらの登場は、GxP の観点からは、リスク、品質保証、およびバリデーションに関して再考する必要がある。

ChatGPT[1]は人工知能 (AI) の利用を次のレベルに引き上げ、既存の AI の枠を超えて世界中のユーザーベースに到達し、Bard[2]に搭載されている LaMDA のような他の大規模言語モデル (Large Language Models ; LLMs) もそれに続いた。しかし、ChatGPT のようなツールを安全性が重要な、製品自体に関連するような環境に適用することを想像できるだろうか？そのためには、ChatGPT とは何かを理解する必要がある：生成型 AI の領域内での大規模言語モデルとして、これは、コンテンツを作成することができるため、機械学習 (Machine Learning ; ML) の他の目的である点推定の予測やカテゴリーまたは分類に関する提案を生成することとは対照的である。とはいえ、この技術は製薬業界の規制分野でも役に立つのだろうか？本稿は品質、リスク、効率の観点からこの問題を検討し、このようなツールがバリデーションと品質保証についての考え方を変えることができるかを議論する。

ChatGPT は一般的に、チャットボットのウェブインタフェースと、その基礎となるモデルを指すために使われている。読みやすくするため、本稿では両者の解釈を同じ意味で使うことにする。

ChatGPT を使用した具体的なケースを説明するセクションでは、次のように表記している：

- 筆者らからの質問や発言は筆者と表示される。
- ChatGPT の応答は ChatGPT と表示される。
- すべての内容は引用符で囲まれている。

これらの ChatGPT の例では、ChatGPT のウェブインターフェースを使用している：  
<https://chat.openai.com/chat> (2023年2月9日記録)。

**筆者：**「ChatGPTは製薬業界の規制分野での使用にも有用か？」

**ChatGPT：**「はい、ChatGPTは製薬業界の規制分野で有用である可能性がある。例えば、医薬品の用途、用法、用量、副作用、他の薬との相互作用に関する情報を提供することができる。ただし、ChatGPTは一般的な言語モデルであり、その出力は医療アドバイスとはみなされないことに注意する必要がある。規制分野では、情報を患者のケアや治療の判断に使用する前に、常に適格な医療専門家によって検証され、確認されるべきである。さらに、規制された分野でのAIの使用は、特定の法律、規制、ガイドラインに従う必要がある。いかなる使用もこれらの要件に準拠していることを確保することが重要である。」

この回答について筆者らは以下の点に注意している。第一に、ChatGPTは有用であると主張していない。ただ、有用である可能性があると述べている。筆者らはこの見解を共有し、ガバナンスと品質保証に関して詳しく説明する。

第二に、ChatGPTは、製薬業界で使用される可能性のある活動やタスクについて、非常に広範でハイレベルな概要を提供している。しかし、それをどのように適用できるかについての詳細情報は提供していない。また、筆者らはChatGPTが同等に重要性を持つさらなる使用事例を提供できると考える。

第三に、筆者らはChatGPTは、自身の回答で述べているように、一般的な言語モデルであり、それに応じて「適格な医療専門家」や「医療アドバイス」といった用語を使用している点に注目している。医薬品の役割の特定の側面（品質保証など）については、ここでは言及していない。筆者らはChatGPTが、その回答は適格な医療専門家によって検証されるべきであると指摘していることを確かに認識する。

まとめると、ChatGPTはこのようなシステム、あるいは他のカスタマイズされた生成型AIシステムが規制分野での使用について、検討するための興味深い出発点を提供している。しかし、最初の例に示されるように、これは情報に対する批判的な思考と注意深い評価の始まりに過ぎない。

本稿の残りの部分は以下のように構成する：LLMの仕組みとChatGPTが構築された概要を提供し、チャット用に最適化された生成型AIの例としてChatGPTの使用事例を説明し、チャットやテキストの作成といった初期のアイデアからソフトウェア開発までの使用事例を紹介する。次に、ChatGPTとLLMの使用に関する一般的なリスクについて説明する。ChatGPTの機能を、医薬品のバリューチェーンに関連した特定の用途に移し、ChatGPTとLLMを一般的に考慮しながら製薬業界の規制分野での使用の背景を詳しく説明する。そして本稿は、LLMの利用が現在の誇大広告を超えてどのように進化する可能性があるかについての展望で締めくくる。

## LLMsの概要とトレーニング

簡単に言えば、言語モデルとは単語の並びに対する確率分布である。つまり、与えられたテキストのシーケンスに対して、モデルは次に何が来るかを予測することができる。大規模言語モデルという用語は明確に定義されていないが、通常は非常に大規模なニューラルネットワークに基づく言語モデルを指す。次の単語を予測するトレーニングを通じて、LLMs は言語の基本的なパターンや構造、および単語に関する事実を学習する。これらのモデルはその後、幅広い言語処理や言語理解タスクに適用することができる。

最初のニューラル言語モデルが提案されたのは 20 年以上前だが[3]、現代の LLMs の優れた能力につながるいくつかの重要な革新がある。特に：

- ハードウェアの進歩：グラフィックプロセッシングユニット (Graphics Processing Units; GPUs) やテンソルプロセッシングユニット (Tensor Processing Units; TPUs) のような特殊なチップの利用可能性が、大規模なデータの高速度処理を可能にする
- ソフトウェアの進歩：新しいニューラルネットワークアーキテクチャ、例えばテキストのシーケンスをより効率的に処理するトランスフォーマーモデル、およびより高度な数値最適化技術
- より大規模で多様なトレーニングデータセット：現代の LLMs は、旧来のモデルよりもはるかに大規模でより多様なデータセットでトレーニングされている

例えば、GPT-3 は[4]、最も強力な言語モデルの一つと考えられており、ChatGPT は約 45 テラバイト (数千億の単語) の多言語テキストデータ (クロールされたウェブサイト、書籍、ウィキペディア) でトレーニングされた、1750 億のパラメータを持つ大規模なトランスフォーマーモデルに基づいている。

このような言語モデルのトレーニングは、しばしば事前トレーニングと呼ばれる。なぜなら、事前トレーニングされた言語モデルは通常、特定のタスクを実行するために教師あり (ラベル付き) データセット上で望ましいタスクに合わせてカスタマイズされているからである：例えば、感情分類や特定のエンティティの抽出などである。リソースのかかる事前トレーニング (一度だけ行う必要がある) は、モデルに一般的な言語能力を与え、一方、カスタマイズはモデルのパラメータを特定のタスクに適応させる。このアプローチは、Generative Pre-trained Transformer の略である GPT という名前にも反映されている。

GPT-3 のような LLMs は、入力テキストと適切に作成されたプロンプト (つまり、タスクを説明する自然言語の指示) をモデルに提供するだけで、追加のカスタマイズなしに自然言語処理 (Natural Language Processing ; NLP) タスクに適用できる。このいわゆるゼロショットアプローチには強い利点がある：カスタマイズステップのためにデータをラベル付けする必要がないことと、同じモデルをさまざまな NLP アプリケーションに使用できることである。しかし、このアプローチでは通常、カスタマイズされたモデルのモデルパフォーマンスに達しない。実際のプロンプトの前にいくつかの入力と出力の例を置くことは、しばしば著しく良い結果につながる。このアプローチはフューショット学習と呼ばれる

る。

したがって、GPT-3のようなLLMsは、NLPタスクを実行するために使用することができる。しかし、これらのモデルは、統計的に最もありそうな次の単語を予測するようにトレーニングされただけで、しばしば指示に従うのが得意ではなく、時には信頼に値しなくかつ有害な出力を生成することもある。ChatGPTはGPT-3からカスタマイズされた派生物であり、LLMをよりユーザーの意図に沿うようにしている。つまり、より役立ち、かつ安全な応答を会話形式で生成することである[5]。

そのために、モデルには人間のラベラーが作成したテキスト入力とそれに対応する出力の例が提供される。さらに、LLMを中心にした比較的新しいアプローチである人間のフィードバックからの強化学習（Reinforcement Learning with Human Feedback ; RLHF）がモデルのトレーニングに使用されており、人間のラベラーによってランク付けされた出力に基づいてモデルを最適化する。ランキングは、出力を書くよりも簡単なタスクで、はるかに効率的である。望ましくない出力に対する追加の保護層として、ChatGPTはアルゴリズムを使用して有害なコンテンツを分類し、フィルタリングしている。

この数年間、大規模テクノロジー企業、AIスタートアップ、研究イニシアティブによって、さまざまなLLMsが開発されてきた。最も人気のあるモデルは、ほぼ間違いなくOpenAIのGPT-3モデルファミリーであり、推論と微調整のためにアプリケーションプログラミングインターフェース（Application Programming Interface ; API）を介した商用利用可能なクローズドモデルである。グーグルの人気のあるPaLMモデルは非常に強力な機能を持っていると言われているが、このクローズドモデルは一般には公開されていない。その他のよく知られたモデルとしては、Meta社のOPT（利用に制限のあるオープンモデル、たとえば非商用利用のみ）や、研究イニシアチブBigScience社のBLOOM（利用にわずかな制限のあるオープンモデル）がある。今後、多くの新しい基本的なLLMsや、専門モデル（対話システムなど）が開発されることが予想される。

ChatGPTがどのように構築され、トレーニングされたのか、この概要を考慮して、本稿では規制分野での応用に及ぼす特定の品質と制御の影響について詳しく説明する。

## ChatGPTの使用

チャットタスクに焦点を当てたLLMアプリケーションの代表的なものとしてChatGPTを使用することで、次のような使用事例を確認した：

- ブレーンストーミングアシスタント：ChatGPTは、さらなる検討ポイントを提供し、アイデアに基づいて提案を洗練させることができる。
- 執筆アシスタント：制作するコンテンツのアイデアを提供すると、ChatGPTは詳細やさらなる情報を追加して完全な段落を作成することができる。また、特定の読者やスタイルに合わせてテキストを調整することもできる。

- 微調整アシスタント：製薬業界における言葉による貢献の多く、例えば論文、評価、およびレポートは、ビジネス上の内容から本質的に複雑である。しかし、これらの出版物で提示されている複雑な文法は、しばしば不必要なものである。ChatGPTは、筆者あるいは読者の立場で、言葉を単純化するために使うことができる。
- ナレッジベース検索：あるトピックについて、簡単な理解からより深い専門的な情報まで、さまざまなトーンで探索および説明できる。冒頭の例で示したように、ChatGPTはチャットの性質により、得られた洞察を洗練させ、批判的な質問と探求のバランスをとることを可能にする。
- ソフトウェア工学：ChatGPTは、意図する使用の説明があれば、シンタックスを含む完全な実装を提供することができる。また、実装を検証するための自動テストを提供することもできる。
- ソフトウェアの校正と頑健化：ChatGPTは実装のエラーを検出し、コード内でそれらを解決するヒントを提供することができる。

#### ケース 1：キーワードセットから段落を作る

**筆者：**「製薬業界の品質保証の観点から、コンピュータ化システムにおけるデータの完全性に関するセクションを 100 単語以内で作成してください。」

**ChatGPT：**「コンピュータ化システムにおけるデータの完全性は、医薬品の品質と安全性を確保するために重要です。製薬業界の品質保証の観点から、データの完全性とは、製薬業務で使用される電子データの正確性、一貫性、信頼性を指します。これには、実験機器、製造プロセス、およびビジネスシステムから生成されるデータが含まれます。品質保証は、データが未承認の変更から保護され、意図された状態を正確に反映することを確保しなければなりません。これは、安全なシステム、データのバックアップ手順、およびアクセス管理などの効果的なコントロールを実装することで達成できます。さらに、データの正確性を検証し、不一致を特定するために、定期的な監査と検査を実施すべきです。データの完全性を確保することは、規制要件を満たし、患者や利害関係者の信頼を維持するために不可欠です。」

**筆者：**「ソフトウェア開発の観点から、コンピュータ化システムにおけるデータの完全性に関するセクションを 100 単語以内で作成してください。」

**ChatGPT：**「製薬ソフトウェア工学の環境では、データの完全性とは、コンピュータ化システムに保存されたデータの正確性、一貫性、信頼性を指します。これは、患者の安全性に直接影響するため、医薬品ソフトウェアの重要な側面です。データの完全性を確保するために、製薬会社は厳格なバリデーションプロセスを実施し、定期的にバックアップを行

い、未承認のアクセスやデータ操作を防ぐための強固なセキュリティ対策を講じなければなりません。さらに、ソフトウェアシステムは、製薬業界における電子記録と電子署名の使用に関するガイドラインを定めた 21 CFR Part 11 などの規制基準に準拠しなければなりません。データの完全性を維持するには、技術的、組織的、手続き的な対策を組み合わせ、データのライフサイクルのすべての段階でデータが確実に保護されるようにする必要があります。」

ChatGPT からのこれらの回答を以下のように評価している。どちらの回答もデータの完全性とその重要性について適切な説明を提供しているが、ALCOA（帰属性、可読性、同時代性、原本性、正確性）の原則のすべての側面をカバーしておらず、また、トレーサビリティのようなよく知られた付加事項も欠けているため、回答は不完全である。どの回答も、データの完全性、製品の品質、および患者の安全性との関連を指摘していない。しかし、これらの回答はデータ保護とアクセス制御の重要性を強調している。

予想通り、どちらの回答も類似している。リクエストのわずかな、しかし緻密な違いが回答によく反映されており、その違いは一目瞭然である。2つ目の回答は、システム設計のいくつかの側面について言及している。しかし、データの完全性を確保するためのソフトウェア開発ライフサイクルのアプローチといった重要な要素については言及されていない。それでも、回答はビジネス事項に焦点を当てており、これは、最初のリクエストの文脈を「覚えている」ことが原因かもしれない。こうして、チャットの履歴によって回答が異なることを考慮することが重要である。驚くべきことに、単語数はどちらの場合も 100 を上回っている（～120）。したがって、明確に策定された非機能要件は満たされない。

## ケース 2：自然言語を洗練させる

筆者：「以下のセクションのテキストを簡略化してください（例文は[6]から引用した）：使用目的、アドレス、許容されるリスクが特定されたら、そのソリューションが特定の品質への期待を満たすように設計されるべきである。モデリング決定のその前の段階の影響を過小評価することはできない。ユーザーが AI の出力に基づいて決定を行うために、ソリューションの結果に直面するあらゆるケースにおいて、ソリューションがオペレーターによって例外のみが処理されるモードで動作する状況と比較して、異なるモデリング戦略が追求される可能性がある。しかし、最初のモデルが評価されれば、使用事例に関するさらなる理解が得られるため、これまでの決定を批判的に反映することが重要である。早期のフィードバックとアラインメントのためのプロトタイプ戦略は、開発プロセスの流れの後半における受容リスクをさらに低減する。この建設的なプロセスにおいて、決定事項に関する文書化は、モデルの正当化や監査の環境における説明可能性メカニズムの選択にとって重要である。」

**ChatGPT:**「ソリューションは、その望ましい品質基準を満たすように作られるべきである。どのように使われるのか、誰が使うのかによって、ソリューションの作り方は変わってくる。早期の選択の影響は無視できない。ユーザーが常にソリューションを使って決定する場合と、例外の時のみオペレーターを助ける場合では、ソリューションの結果は違ってくる。選択したことを評価し続けることが重要である。早期にソリューションをテストすることは、リスクを減らすのに役立つ。決定したことを記録しておくことは、将来、特定の選択をした理由を説明するのに役立つ。」

ChatGPT からのこれらの回答を以下のように評価している。テキストはかなり簡略化されている。しかし、セクションの内部ロジックはもはや反映されていない。例えば、「決定事項」が本文に十分に組み込まれていない。ChatGPT は AI をキーメッセージの一部として認識していない。したがって、文脈の一部が欠落し、新しいテキストの誤解を招くことになる。さらに、「説明」の役割が逆転しているように思われる。筆者らは ChatGPT の回答を、プロセスの結果が行われた決定を説明すべきであるという意味で解釈しているが、元の文章はアルゴリズムの説明可能性について詳しく述べていた。しかし、筆者らはこの段落を筆者の視点から見ており、複雑なトピックを扱う技術論文から抜粋されたこの孤立した部分の意味を理解するのは、人間には難しいかもしれない。

### ケース 3：ソフトウェアまたは基本ソフトウェアキットを作る

**筆者:**「表形式のデータセットで線形回帰を実行できる Java コードを提供してください。また、このクラスのテストケースも提供してくれますか？」

回答として、筆者らは要求された Java コードと、最後にある短い説明文を受け取った。読みやすさのためにここには掲載していないが、コピーは保管されている。

筆者らは以下のように評価している。コードが Java 17 でコンパイルおよび実行できることを検証した。コードは一般的に読みやすい：線形回帰の一般的な変数名が使われている。コードスタイルはベストプラクティスに準拠しており、例えば、実際の推定アルゴリズム用のコンストラクタを使用し、別の予測メソッドを提供している。テストケースは、コードの妥当な応用例ではあるが、ソフトウェア工学的な意味でテストされていない。つまり、機能性の検証は、予想される予測と実際の予測を比較することによって行われる。そこで、線形回帰の方法論的な質は評価されるが、実装の正確性は評価されない。

## LLMs の使用に固有のリスク

ChatGPT は様々な分野で能力を持っているが、前述したようにリスクを伴わないわけではない。これをあらゆる使用環境において考慮する必要があるが、規制の厳しい製薬業界ではリスク管理の観点からなおさらである。筆者らは、LLMs の一般的な使用にも当てはま

る以下 10 個の重要なリスク項目を特定した。

### 品質と正確性

ChatGPT は迅速にコンテンツを提供できる。しかし、すでに例で見たように、コンテンツは対象分野の専門家によってレビューされ、検証される必要がある。LLM は内容を創造する（「幻覚を引き起こす」）可能性があるため、どの部分が意図する利用にとって価値があるかを判断するのはユーザーの責任である。

### 参照情報と検証

ChatGPT は、情報を確認するための参考文献自体を提供していない。チャットの機能性を使って、このトピックに関する参考文献を引き出すことができる。しかし、これらの「参考文献」は質が低いか、または捏造されている可能性がある。LLM は参考文献の典型的なパターンを学習したため、タイトルや著者を創造するかもしれないが、参考文献の完全性を検証する方法を知らない。これは、生成された回答の一般的なトレーサビリティと信頼性に懸念を引き起こす。また、各回答は対象分野の専門家によって注意深く検証されなければならないという筆者らの主張を強めるものである。しかし、筆者らの経験からでは、チャット機能を使って結果の説明を求めると、時に興味深い背景情報が明らかになることがある。

### 再現性

ChatGPT による回答は変わる可能性がある。すなわち、「回答を再生する」オプションが意図的に提供されている。したがって、結果と内容のトレーサビリティは限定的であり、また前述した例のように、入力に対する感度はかなり高い。トレーニングに使用されたハイパーパラメータ構成の再トレーニングと更新は、ML モデルの他の応用と同様に、回答の他のばらつきの要因を含む可能性がある。つまり、モデルは時間とともに進化している。

### 最新情報

トレーニングデータに関して凍結モードで動作している場合、言語モデルは特定の時間範囲までの情報しか取り込まない（執筆時点では、ChatGPT のトレーニングデータは 2021 年第 4 四半期までをカバーしている[7]）。この時間軸を超える情報は、モデルの特定のバージョンでは知ることができない（あるいは限られた量しか知ることができない）ため、現在の状態の情報を反映しない回答が導き出される可能性がある。

### 知的財産権と著作権

ある意味、トレーニングデータユニバースの全体が各回答のソースと考えることができる。よって、正確な情報（コード）はすでに第三者によって作成されているかもしれない。

したがって、結果の知的財産権を二重確認することは、法的リスクを軽減するために重要である。また、ChatGPT に提供された入力も、知的財産権と著作権の考慮事項に該当する可能性があるため、影響を受ける。

### バイアス

バイアスは、GPT-3 モデルでの事前トレーニング、微調整、モデレーションなど、ChatGPT のトレーニングステップのさまざまなステップから生じる可能性がある。事前トレーニングは教師なしデータで次の単語を予測するためにトレーニングされたため、モデルはこのデータセットからバイアスを学習する。微調整のプロセスで、オペレーターはチャットの入力に対して「ゴールデンレスポンス」を提供した。たとえ OpenAI の原則の基準で行われたとしても、この人為的な意見からのバイアスは排除できない。最終的に ChatGPT は有害なコンテンツを防ぐためにモデレートされるが、これもまたラベル付けされたデータでトレーニングされるため、バイアスがかかる可能性がある。まとめると、この 3 段階のプロセスにおける複合的な効果は、ユーザーの視点からは検証できない。そのため、ユーザーは批判的思考せずに ChatGPT の結果を活用すると、自身のコンテンツを歪めてしまうリスクがある。

### 意見の形成

前述の点に関して、特にインシデントの評価などの評価演習では、ChatGPT が「意見」を形成し、したがって意思決定プロセスに影響を与える可能性がある。しかし、注意したいのは、人間の意思決定にもバイアスがないわけではない。こうして、ChatGPT は手近な質問に対する第二の観点を提供するか、またはあるトピックについて異なる角度から質問された場合は複数の観点を提供することがある。

### サービスの可用性

この執筆時点では、ChatGPT の無料版は重いユーザー負荷により現在安定していない。しかし、ChatGPT はプロフェッショナルな環境で使用するためのプロバージョンと API アクセスも提供している。

### データの機密性

ChatGPT インタフェースに入力されたすべての入力は、ユーザーのコントロールを超えている。したがって、法的リスクを軽減するために、内部情報や機密情報を送信すべきではない。このため、特に製薬業界およびヘルスケア分野では、個人を特定できる情報やビジネスクリティカルな情報を含む使用事例が多いため、ChatGPT の利用が制限される。

### 規制とガバナンス

規制上または内部ガバナンス上の理由により、生成的 AI ツールの使用が使用事例で許可されているかどうかを確認する必要がある。これらのリスクのため、一部の企業はすでに ChatGPT や類似のシステムをネットワーク上でブロックしている。しかし、これが解決策だろうか？ 次のセクションでは、ChatGPT を LLMs の代表として、製薬業界をサポートできる具体的な使用事例を詳しく紹介する。

### 製薬業界の規制分野における ChatGPT の使用事例

一般的な応用分野とリスクを考慮すると、ChatGPT のような LLMs が製薬業界、ひいては患者に価値を提供できる可能性はどこにあるのだろうか？

製薬のバリューチェーンに沿って、筆者らは 5 つの使用事例があると考えている。これらについては以下のように議論され、また図 1 と表 1 に示されている：

1. 研究：ChatGPT は、特定のターゲットメカニズムに関するアイデアを生み出すブレインストーミングアシスタントとして使うことができる。
2. 開発：ChatGPT は、症例の説明が提供される場合、研究レポートの作成を支援するために使うことができる。
3. 市場アクセスと商業化：ChatGPT は、患者がより理解しやすいように医薬品の説明を簡略化するために使うことができる。
4. 製造：ChatGPT は、生産プロセスを管理するために必要なソフトウェアの一部を作成するために適用することができる。
5. 流通と市販後：ChatGPT は、ソーシャルメディアのような非古典的な医薬品安全性監視ソースからの洞察を要約するために適用することができる。

表 1 では、これらの状況で ChatGPT のような LLM の使用に対する評価を示す。使用事例に基づくと、LLM の使用はプロセスの理解や対象分野の専門知識を置き換えるものではないと考える。さらに言えば、これらの技術の応用には、LLM が提供する結果を検証するための使用事例と内部メカニズムに対する強い理解が必要で、またこれらのモデルのブラックボックス的性質から生じるさらなるリスクに対する感度も必要である。

したがって、補助的なツールの位置付けが最も有望な用途があると考えられる。技術的なコンテンツ（コード、自動テスト）であれ、専門的なコンテンツ（作成されたテキスト、レポート、セクションなど）であれ、アウトプットを評価するためには批判的思考が求められる。批判的思考を促進し、品質を維持するためには、標準作業手順書によって、出力の使用と洗練方法を定義する必要があり、また LLM からの援助が倫理的である分野とそうでない分野を定義する必要もある。バイアスのリスクが重要であると考えられる例のひとつは、評価やアセスメントに関するもので、これはそもそも対象分野の専門家の見解に影響を与えたり、シフトさせたりする可能性がある。

## 規制のある製薬業界における LLMs の幅広い応用

ChatGPT やその他の言語ツールは、重要なタスクのための補助的ツール以上の潜在能力があるだろうか？筆者らの意見では、オープンで公開されたモデルを GxP 規制の環境に適用することは、以下の理由から困難であろう：

- 企業は、モデルの開発プロセスを直接管理することも、トレーサビリティを確保することもできない。
- 企業は、知の地平、つまり、関連するトレーニングデータがトレーニングセットに入った最後の時点を管理することができない。
- 企業は、トレーニングプロセスに使用されたラベルやフィードバックの管理が不十分である。

しかし、ChatGPT は NLP の分野で AI の力を実証した。すなわち、管理された開発環境では[8]、製薬業界またはライフサイエンス分野向けに特別にトレーニングされたソリューションの開発が有望であることを意味する。もしこのような LLM が、対象分野の専門家によって管理されたプロセスに統合され、バリデーション演習で人間・AI チーム[9]のパフォーマンスが改善されたという証拠が提供された場合、これらのアプローチはプロセスに価値を与え、最終的には製品品質に価値を与えることができる。AI 成熟度モデルの観点からは[10]、これは検証レベル 3 のソリューションを指す。

ガバナンスと管理の観点から、知識体系の再トレーニングと更新に主眼を置きながら、リスクはアプリケーションのライフサイクルにおける 5 つの品質の次元[11]に沿って監視される必要がある：

1. 予測力：出力は十分な品質で、ユーザーの質問に答えているだろうか？
2. 較正：言語モデルは、使用事例の関連するすべての次元において十分に優れた性能を発揮するか、あるいは一部の使用事例において重大なバイアスを示すか？
3. 頑健性：出力は時間とともに十分に安定し、出力の変化は理解可能だろうか？
4. データ品質：トレーニングに使用されるデータ、およびシステムへの入力として提供されるデータは、データガバナンスのベストプラクティスに従って品質が保証されているか(例：データのラベリング、トレーニングケースとプロダクションケースの代表性の分析)？
5. 使用試験：ユーザーは手近な問題と言語モデルの結果の関連性を理解しているか？言語モデルの出力に対して、質問し、検証し、応答を修正することによって、適切に反応しているか？

このような管理された状況であっても、GxP 上の重要な環境では、出力を検証するために人間によるレビューが必要であると予想される。筆者らが見る限り、製薬業界やライフサイエンス全般における声明が真実か虚偽かを真に評価できるのは、その対象分野の専門

家だけである。

## 結論

LLMs と ChatGPT は、これからも存在し続ける。この技術の使い方を学ぶのは、ユーザーであり、対象分野の専門家である我々の責任である。この技術が日常的なビジネスアプリケーションでより容易に利用できるようになれば、その重要性はさらに高まる。批判的な思考と組み合わせることで、これらの LLM の成果は、製薬業界の規制された環境においても、単独のソリューションとしてではなく、作業ツールとして、ソフトウェア工学から規制の文書化まで、さまざまな業務部門の効率を高めることができる。ウェブインターフェイス、ビジネスアプリケーション、API、あるいは商用のクラウドサービスを介してこれらのサービスを利用するには、厳格さ、品質保証、および法的リスクの軽減を前提条件としなければならない。適切な使用事例を見つけるにはクリエイティブでなければならないかもしれない。

さらに、より専門的な LLM が人気を集めるだろう。他の安全性が重要で、GxP 管理下の分野と同様に、これらのモデルも管理された品質とベストプラクティスの条件下で開発されなければならない。この厳密さには、入力データの品質保証、開発プロセスとバリデーションの管理、生産的なモニタリングと品質リスク管理プロセスが含まれる。これはまた、真の人間-AI チームのコラボレーションに 2 つの新たな強力な次元、すなわち、AI のそれぞれの強みと、適切な目標制御モデルを持つ認知知能に焦点を当てて効率を向上させることを解き放つ。

従って、企業は適切な基準と管理を設定することで、これらの発展に対応することが最も重要である。この記事では、そのような LLM を使用する際に固有の潜在的な使用事例とリスクについての一般的な概要を提供した。このガイダンスに基づき、このようなモデルの応用は常に、対象分野の専門家、開発者、およびさらなるスタッフのための作業補助の役割として、あるいはコンピュータ化システム自体の一部として、具体的な使用目的の下で評価されなければならない。

本文以上

<図表の説明>

図 1 バリューチェーン全体における人工知能の応用

表 1 製薬バリューチェーンに沿った選択された LLM アプリケーションの評価