大規模言語モデルの業務活用における制約と対応方法

昨今、大規模言語モデル（LLM）の利用が広がりを見せる中、金融機関において業務への活用検討・実証実験が進んできている。LLMの業務活用を検討する際は、ハルシネーションをはじめとした技術的制約や動作の特徴を正確に把握した上でユースケースや構築方法を検討する必要がある。また、チューニングにも配慮する必要がある。

LLMの業務活用を検討する際に考慮すべき点

現在、大規模言語モデル（LLM）を金融業務に活用しようという動きが多く見られる。しかしながら、LLMは当然ではあるが万能なAI技術ではない。動作の特徴を理解し、業務への活用にあたってはその制約を考慮して活用を検討することが必要となる。自社内のデータとの連携や生成した文章の評価が不可欠であり、データオーナーである金融機関自身が対応すべき範囲も広い。

金融機関が当初、懸念していたのはデータ保護であった。例えば、入力したデータをLLMの再学習に使われてしまう懸念や、データセンターが海外に位置することによるデータ越境移転の対応の懸念といったものがこれにあたる。しかし、MicrosoftがAzure OpenAI Serviceの一般提供を開始し、2023年7月には日本リージョンでのサービス提供も発表されたことにより、データ保護への対応が行いやすくなり、当初の懸念は大幅に後退した。この結果、LLMの導入が加速したといえよう。現在では、次のフェーズとして、自社データと組み合わせ、どのような業務に活用できるか各社で具体的な検討を始めている[1)]。

以下、検討を進めるにあたり、金融機関のトップマネジメントにとって参考となる基本的で代表的な制約と対処方法について示すこととしたい。

まず多くの人が想像するものは「ハルシネーション」であろう。ハルシネーションはLLMが事実に基づかない「嘘」を生成してしまう事象である[2)]。加えて、学習時点までのデータに基づく文章生成を行うため、最新の情報に基づく文章を生成するためには十分な参考情報を与える必要がある点も、正確性の確保という観点からは注意が必要である。

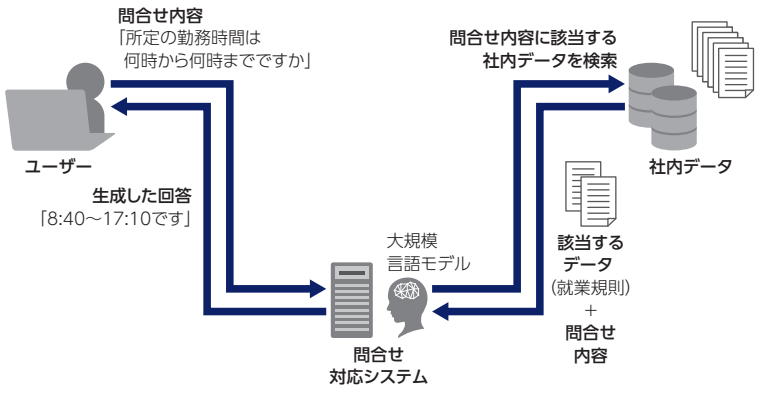
これらの制約は、プロンプトを介しデータを動的に絞り込むGroundingという手法などによって抑制することが考えられる（図表1）。また、生成した文章に対し必ず人間が正確性を確認することも有効である。

次に、多くの計算リソースを消費することもLLMの制約として考慮に入れる必要がある。LLMは学習時だけでなく動作時にも比較的多くの計算リソースが必要であり、運用コストが大きくなる恐れがある。

これは、適切な規模（パラメータ数）のLLMを使用したり、一部のタスクを別の方法で処理するといった方法により抑制することができる。また、あえて無理にLLMを利用しないというのも選択肢になろう。

その他の制約として言語格差、計算や最適化などの確立された解法に基づき答えが導出されるタスク等についてはLLMが必ずしも得意とする分野ではないことも考慮

図表1　照会対応におけるGroundingのイメージ



(出所) 野村総合研究所

NOTE

1) 各社の状況は報道及びニュースリリースより。一部推測を含む。2023年8月末時点。
2) LLMは確率に基づく文章を生成するモデルであり、文脈に適合する文章を生成しているに過ぎないため発生する。

する必要がある。これらの制約はユーザーとの入出力に直接関係しないプロンプトを英語やプログラミング言語での表記とすることや、LLMが得意としないタスクは別の方法で処理するといった方法で抑制できるだろう。

適切なユースケース選定に加えチューニングも重要

各金融機関のニュースリリースや報道等によると、現時点では主に次の5つに分類されるユースケースで業務への活用を検討していると思われる。照会対応（社内問合せ対応など）、文章生成・要約（稟議書や契約書素案作成など）、情報収集・翻訳、アイデア創出（アイデアの壁打ちなど）、プログラミングの5つである。いずれのユースケースにおいても制約を適切に把握し対処する前提で、現時点のLLMのケイパビリティを踏まえ業務に活用できる可能性は十分にあり、検討は引き続き進んでいくものと思われる。

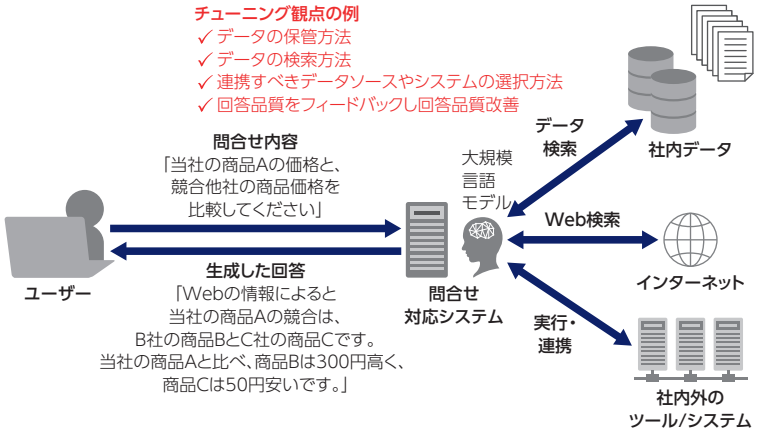
その際、チューニングと継続的な改善も併せて進めていく必要がある。例えば代表的なユースケースである照会対応を自社データも活用し実現する場合、LLMの回答精度を高めるためのチューニングが必須である（図表2）。自社データを連係する際、自社データをどのような形式で保管し、どのようにユーザーの問合せに対応した適切なデータを検索しLLMに渡すべきか、環境や実際に取扱うデータによって大きく異なってくる。

さらに、指示に基づき複雑かつ正確な情報をLLMに処理させ回答となる文章を生成する場合は、自社データだけでなく他のツール等との連係も必要になるだろう（図表2）。その場合、問合せ内容からどのツールを正しい順序で実行すればよいかを判定する必要があり、プロンプトはさらに複雑化する。

このため、それらのプロンプトエンジニアリングを支援するライブラリを活用する必要があるかもしれない。プロンプトエンジニアリングによって回答精度が大きく変わる可能性があり、チューニングの際のひとつのテーマである。

また、LLMが生成した文章（回答）が良い文章なのか悪い文章なのかをユーザー自身が判定しフィードバックすることで、よい回答をユーザーに提示するように継続的なチューニングを行うことも重要となってこよう。

図表2　照会対応におけるチューニングのイメージ



（出所）野村総合研究所

Writer's Profile

金子 洋平 Yohei Kaneko
金融デジタルビジネスリサーチ部
シニアリサーチャー
専門はリテール金融、金融DX
focus@nri.co.jp