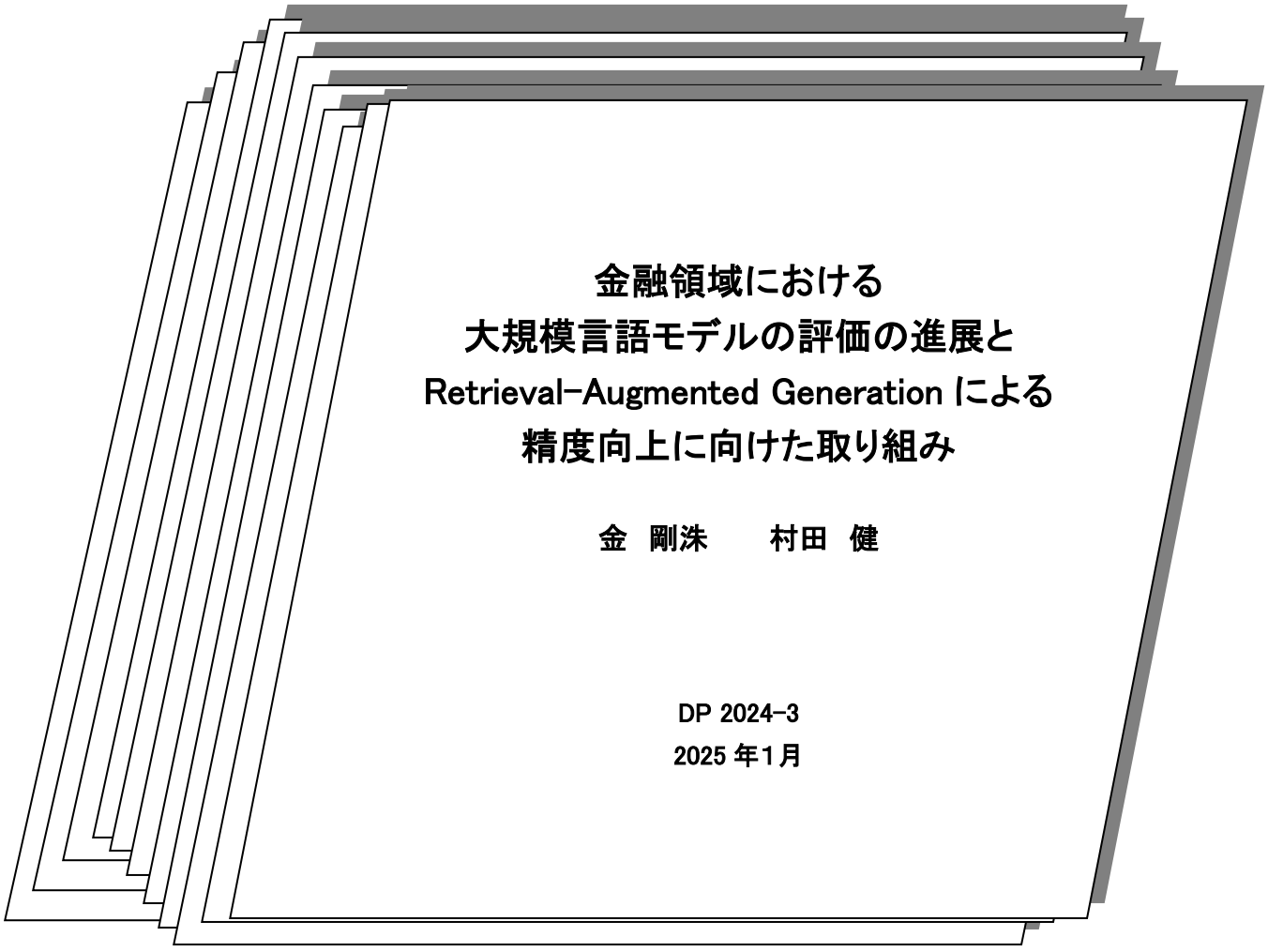
FSA Institute
Discussion Paper Series



金融領域における
大規模言語モデルの評価の進展と
Retrieval-Augmented Generation による
精度向上に向けた取り組み

金　剛洙　　村田　健

DP 2024-3
2025 年1月

金融庁金融研究センター
Financial Research Center (FSA Institute)
Financial Services Agency
Government of Japan

金融庁金融研究センターが刊行している論文等はホームページからダウンロードできます。
https://www.fsa.go.jp/frtc/index.html

本ディスカッションペーパーの内容や意見は、全て執筆者の個人的見解であり、金融庁あるいは金融研究センターの公式見解を示すものではありません。

＜金融庁金融研究センター　ディスカッションペーパー　DP2024-3（2025 年 1 月）＞

金融領域における大規模言語モデルの評価の進展と Retrieval-Augmented Generation による精度向上に向けた取り組み

金　剛洙*　　村田　健*

概　　要

2022 年の ChatGPT 登場以降、生成 AI、特に大規模言語モデル（Large Language Model、LLM）への注目が急速に高まっている。金融機関においても、業務効率化や顧客対応の高度化に向けて、これらの技術の活用検討が進んでいる。一方で、LLM は確率的なモデルであり、その出力の正確性が必ずしも担保されるわけではない。学習データ自体に事実誤認や偏見が含まれている場合、誤った情報やバイアスをもとにもっともらしいテキストが生成されてしまう可能性がある。また、モデル内部の推論プロセスが不透明で説明性が乏しい点も懸念材料である。こうした課題は、顧客との信頼関係が重要な金融機関にとって、技術活用にあたって慎重に検討すべき問題といえる。金融規制当局も、生成 AI/LLM が金融セクターに与える潜在的な機会とリスクに大きな関心を示している。金融安定理事会（Financial Stability Board、FSB）は、生成 AI を含む AI の急速な進展と金融セクターにおける AI の利用拡大を踏まえ、2017 年 11 月の AI 報告書を更新する形で、AI が金融安定に及ぼす潜在的な影響に関する報告書を 2024 年 11 月に公表した。同報告書では、生成 AI により文書要約、情報検索、コード生成などの新たなユースケースが登場していることを指摘しつつ、サードパーティへの依存やサイバーセキュリティ、モデルリスク、データ品質、ガバナンスといったシステミック・リスクを増大させる可能性のある AI 関連の潜在的な脆弱性を特定している。

そこで、本ペーパーは、文献調査と実証分析を通じて、金融領域における LLM のユースケースを明らかにし、現状の課題を抽出するとともに、それらの課題に対する技術的解決策や今後の発展の可能性を検討することを目的とする。特に、金融領域における LLM の評価手法の最新の研究の状況と Retrieval-Augmented Generation（RAG）の性能、及び性能向上に向けた取り組みに焦点を当てる。金融分野において LLM を導入する際、特に慎重に検討されるべきはモデルの出力をどのように評価するかという点である。一般的な言語モデルの評価は、主に正確性や生成内容の一貫性、ハルシネーション（誤った情報の生成）の頻度といった性能指標に基

*東京大学大学院工学系研究科 学術専門職員（金融庁金融研究センター特別研究員）
本稿の執筆に当たっては、金融庁牛田遼介様、伴ちひろ様をはじめ多くの方々に有益な御意見をいただいた。
なお、本稿は、筆者らの個人的な見解であり、金融庁及び金融研究センターの公式見解ではない。

づくが、金融領域に特化した評価指標はさらに詳細かつ多面的である必要がある。金融機関の AI 担当者及び AI スタートアップを含む AI 開発者・研究者ら複数名への有識者インタビューと先行研究の調査より、特に金融領域における意思決定やリスク管理は、より複雑な認知能力が要求され、既存の LLM 評価基準では十分にカバーできない領域が多いことが明らかとなった。また、金融領域の LLM 活用の事例として注目されている RAG のシステムについて、技術的な概要から、実際に導入する上で工夫するべきポイントの整理を実施した。金融庁のガイドライン文書を対象にした実証実験を行い、RAG システムの構築の事前準備として既存の文書を整理するという取り組みにおいても LLM が活用できる可能性が示唆された。

キーワード：大規模言語モデル（LLM）、RAG、説明可能性

１．はじめに

近年、大規模言語モデル（LLM）は急速に発展を遂げており、その一例として代表的なモデルは 2022 年 11 月に OpenAI 社により公開された ChatGPT[1)]である。ChatGPT はその汎用性と高度な言語理解・生成能力を特徴とし、公開から 2 か月で月間アクティブユーザー数が 1 億人を超え、ソフトウェアアプリケーションとして史上最速でのユーザー増加を記録した。LLM は、文章生成、翻訳、コードの自動生成など多岐にわたるタスクで高い性能を発揮し、従来人間にしか対応できないと考えられていた創造的なタスクにおいても人間の能力に匹敵、あるいは凌駕する性能を実現するに至っている。特に、2022 年以降 LLM は急速な進化を遂げており、ChatGPT の登場から約 2 年の間に、GPT-4[2)]、Claude[3)]、Gemini[4)]、Llama3[5)]といった高性能なモデルが次々とリリースされ、その能力は着実に向上している。特筆すべきは、これらのモデルが単なる言語処理にとどまらず、画像、音声、動画といった複数のモーダルを統合的に処理できる能力を獲得しつつあることである。例えば、GPT-4o[6)]や Claude3.5、Gemini などは、画像理解と言語処理を組み合わせた高度なタスクを実行できる。この発展により、金融分野においても、チャートやグラフの分析、マニュアルの図表の理解、文書のスキャンと解析など、より広範な応用が可能になってきており、金融分野にも大きな変革をもたらす可能性がある。また、BloombergGPT[7)]や FinLLAMA[8)]、FinGPT[9)]など、金融特化型 LLM の開発も進んでおり、この分野における研究開発が急速に進んできていると同時に、大手金融機関を中心に、LLM の活用は、社内利用と顧客向けサービスの両面で広がりを見せている。特に海外の大手金融機関を中心に早期から、LLM の開発及び活用が行われている。下記に代表的な事例を掲載する。

- Bloomberg: 金融特化型 LLM である BloombergGPT の開発。幅広い金融データと汎用のデータを合わせてトレーニングしており、金融領域における多様な自然言語処理に対応。
- Morgan Stanley[10)]: 世界中の企業や資本市場を分析し、ファイナンシャルアドバイザーのサポートツールとして活用。OpenAI と提携し、世界中の企業、セクター、資産クラス、資本市場、地域に関する分析を AI が実施。

---

1) OpenAI (2022) Introducing ChatGPT https://openai.com/index/chatgpt/
2) OpenAI (2023) GPT-4 https://openai.com/index/gpt-4-research/
3) Anthropic (2024) Claude 3.5 Sonnet https://www.anthropic.com/news/claude-3-5-sonnet
4) Google (2024) 次世代モデル、Gemini 1.5 を発表 https://blog.google/intl/ja-jp/company-news/technology/gemini-model-february-2024-jp/
5) Meta (2024) Meta Llama 3 の紹介：これまでで最も高性能でオープンな大規模言語モデル https://about.fb.com/ja/news/2024/04/meta-ai-assistant-built-with-llama-3/
6) OpenAI (2024) Hello GPT-4o https://openai.com/index/hello-gpt-4o/
7) Bloomberg (2023) ブルームバーグ GPT のご紹介 ー金融機関向けにゼロから構築された 500 億パラメータを持つ ブルームバーグの大規模言語モデル https://about.bloomberg.co.jp/blog/press-bloomberggpt-50-billion-parameter-llm-tuned-finance/
8) Thanos, Konstantinidis., Giorgos, Iacovides., Mingxue, Xu., Tony G. Constantinides., Danilo, Mandic. (2024), "FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications", arXiv:2403.12285
9) Hongyang, Yang., Xiao-Yang, Liu., Christina,D. Wang. (2023), "FinGPT: Open-Source Financial Large Language Models", FinLLM at IJCAI 2023, Available at SSRN: https://ssrn.com/abstract=4489826
10) Morgan Stanley (2024) AI @ Morgan Stanley https://www.morganstanley.com/press-releases/ai-at-morgan-

- JP Morgan Chase[11]: 社内 LLM 基盤となる LLM Suite をリリース。資産管理部門の生産性向上を目的に、ライティングアシスタンス、アイデア生成、文書要約など金融特化型 LLM として活用を推進。

国内の金融機関も近年の LLM 技術の進展を受け、LLM の導入・活用に関する取り組みを加速させている。国内では、業務効率化や高度な顧客サービスの実現を目的に、LLM を活用したさまざまな事例が見られるが、一部の大手金融機関、大手のベンダーらは LLM の活用だけではなく、金融特化型 LLM の開発にも着手し始めている。

- 三菱 UFJ 銀行[12]：内製化 ChatGPT『AI-bow（アイボウ）』を国内全行員向けにリリース。Chat 機能や RAG の機能を全行に提供しており、検索、文章生成、要約、翻訳、校正、アイデア創出、コード生成など各人の業務のあらゆるシーン・さまざまな用途で利用。
- 東京海上日動[13]：『One-AI for Tokio Marine』を内製開発し、2023 年 10 月から東京海上グループ内で展開。文書のたたき台作成や、各種アイデア出し、検討の壁打ち、営業でのロールプレイ作成など、さまざまな場面で活用。
- 野村総合研究所[14]：金融機関向けに、金融特化型 LLM を高度なセキュリティ環境で使用できる AI プラットフォームの提供を予定。
- 三菱 UFJ 銀行[15]：KDDI と連携し、金融特化型 LLM の開発を発表。KDDI 傘下の ELYZA と、MUFG が出資する SakanaAI も本プロジェクトに参加し、金融業界向けの高度なサービス提供を目指す。

しかしながら、LLM の進化とともに、その評価方法や実用化に向けた課題も浮き彫りになっている。Language Model Evaluation Harness[16]などの汎用的なベンチマーク計測プラットフォームが提案される一方で、現時点では、日本語の金融分野に特化したベンチマークは乏しい状況である。そこで、本ペーパーにおいては、金融分野における LLM の活用動向を明らかにし、現状の課題を抽出するとともに、それらの課題に対する技術的解決策や今後の発展可能性を探ることを目的とする。また、実際の業務へ LLM の導入を検討する際に、RAG のシステムを開発

stanley-debrief-launch

11) JP Morgan Chase (2024) JPMorgan Chase Leads AI Revolution In Finance With Launch Of LLM Suite https://www.forbes.com/sites/janakirammsv/2024/07/30/jpmorgan-chase-leads-ai-revolution-in-finance-with-launch-of-llm-suite/

12) 三菱 UFJ 銀行 (2024) 生成 AI を相棒のように使いこなせ！三菱グループ各社の社内活用事例、最前線 https://www.mitsubishi.com/ja/profile/csr/mpac/monthly/special_feature/2024/06/1.html

13) 東京海上日動 (2023) 全社員向け生成 AI "One-AI for Tokio Marine"の活用開始～ChatGPT による業務効率化を実現～ https://www.tokiomarine-nichido.co.jp/company/release/pdf/231012_01.pdf

14) 野村総合研究所 (2024) 金融機関向けに各社専用の安全・安心な AI プラットフォームを 2025 年度上期提供 https://www.nri.com/jp/news/newsrelease/lst/2024/cc/0910_1

15) 三菱 UFJ 銀行 (2024) KDDI と MUFG が協業強化 https://xtech.nikkei.com/atcl/nxt/news/24/01816/

16) Stella Biderman., Hailey Schoelkopf., Lintang Sutawika., Leo Gao., Jonathan Tow., Baber Abbasil. *et al*. (2024) "Lessons from the Trenches on Reproducible Evaluation of Language Models", arXiv:2405.14782

し、社内専用の対話アプリや検索システムとして活用するというケースは多く存在する。一方で検索の性能、応答の性能が不十分であり、十分な活用がされないというケースも多く、このような課題に対してどのような対応が有効か、どのような工夫が求められるか、というようなRAG システム設計に関する検証を行う。

本ペーパーの構成は次の通りである。第 2 章では LLM の技術的背景を概観し、3 章で金融分野特有の課題と対策を提示し、金融分野における LLM の活用動向を分析する。第 4 章では LLM の評価手法と課題を詳細に検討し、第 5 章で RAG の概要と性能について論じる。第 6 章では LLM と RAG の性能向上に向けた取り組みを探り、最後に第 7 章で本ペーパーの結論と今後の課題をまとめる。

２．LLM の概要

２．１大規模言語モデル（LLM）の定義と特徴

大規模言語モデル（LLM）は、膨大な量のテキストデータを学習し、人間のような自然言語の理解と生成が可能となった人工知能モデルである。2017 年に登場した Transformer アーキテクチャ[17]を基盤とし、BERT[18]、GPT-4、ChatGPT などの革新的なモデルが開発されてきた。LLM の主な特徴は以下の通りである。

- 大規模なデータセットでの事前学習：インターネット上の大量のテキストデータを用いて学習を行う。
- 文脈理解と生成の高い能力：長文の文脈を理解し、適切な応答や続きを生成できる。
- 多様なタスクへの適応性：翻訳、要約、質問応答など、様々な言語タスクに対応可能。
- ゼロショット学習や少数ショット学習の能力：特定のタスクに対する追加学習なしで、または少量の例示のみで新しいタスクを実行できる。
- スケーラビリティ：モデルサイズとデータ量の増加に伴い、性能が向上する傾向がある。

２．２ LLM のマルチモーダル化について

近年の LLM は、テキストのみならず、画像、音声、動画などの複数のモダリティを統合的に処理できるマルチモーダルモデルへと進化を遂げている。このマルチモーダル化により、LLM は文書の理解・生成に加えて、画像の認識や解釈、さらには画像とテキストを組み合わせた複合的なタスクにも対応できるようになった。これらのモデルは単に異なる種類のデータを個別に処理するだけでなく、モダリティ間の関係性を理解し、統合的な分析や推論を行うことができる。例えば、グラフや図表を含む文書を解析する際、視覚情報とテキスト情報を組み合わせ

17) Ashish Vaswani.,Noam Shazeer., Niki Parmar., Jakob Uszkoreit., Llion Jones., Aidan N. Gomez., Lukasz Kaiser., Illia Polosukhin. (2017) “Attention Is All You Need”, *Advances in Neural Information Processing Systems,* 2017-Decem, 5999-6009.

18) Jacob Devlin., Ming-Wei Chang., Kenton Lee., Kristina Toutanova. (2018) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of NAACL-HLT 2019,* pp. 4171-4186

て、より正確な理解と解釈を実現している。このようなマルチモーダル化の進展により、LLM の応用可能性は大きく広がっている。特に、実世界のビジネス環境では、様々な形式の情報が混在しているため、複数のモダリティを統合的に処理できる能力は極めて重要である。今後も技術の発展に伴い、さらなる性能向上と応用範囲の拡大が期待される。

２.３ ChatGPT 以降の特化型モデルの進展

ChatGPT の登場以降、LLM の研究開発は急速に進展し、様々な特化型モデルが開発されている。これらのモデルは、特定のドメインや言語に焦点を当て、より高度な専門性を持つことを目指している。

２.３.１ ドメイン特化型モデル

特定の分野に特化した LLM の開発が進んでおり、以下のような例がある。

- 金融分野：FinGPT
- 医療分野：Med-Gemini[19)]
- プログラミング分野：Code Llama[20)]

これらのモデルは、それぞれの分野の専門用語や知識構造を学習することで、当該分野において汎用 LLM よりも高い性能を発揮することが報告されている。

医療分野では、患者データの機密性や厳格な規制に加え、高いレベルでの専門性が求められることから、AI の活用が困難とされてきた。しかし、Med-Gemini の事例は特化型 LLM の新たな可能性を示している。Med-Gemini は Google の Gemini モデルを基盤に開発された、医療に特化したマルチモーダル AI モデルである。テキスト、画像、電子カルテなどの長文と医療に特化したマルチモーダルなデータを統合的に処理できるような工夫をしており、診断支援、記録要約、紹介状作成、教育など幅広いタスクに対応可能であり、医療記録要約や紹介状作成のタスクで、モデルが専門家と同等以上の評価を達成した。このように、AI 導入のハードルが高い医療分野においても、ドメイン特化型 LLM は汎用 LLM 研究で発展した技術に加えて、専門分野を効率的に学習させる取り組みを行ったことで、当該分野において汎用 LLM 以上のパフォーマンスを発揮している。このような成功事例は、金融分野を含む他の高度な専門分野にも応用可能であり、LLM の可能性をさらに広げることが期待される。

２.３.２ ドメイン特化の各種手法について

LLM は、幅広い用途に対応する汎用的な言語処理能力を有するが、特定の領域における精度向上や応答品質の向上を目的として、ドメイン特化が必要とされるケースが多い。本節では、LLM を特定領域に特化させるための代表的な手法について概説する。なお、実務の現場におい

19) Google (2024) Advancing medical AI with Med-Gemini https://research.google/blog/advancing-medical-ai-with-med-gemini/
20) Meta (2023) Introducing Code Llama, a state-of-the-art large language model for coding https://ai.meta.com/blog/code-llama-large-language-model-coding/

ては、これらの代表的な手法を組み合わせる形にて、LLM が活用されている。

（1）プロンプトエンジニアリング

プロンプトエンジニアリングは、モデルに入力するプロンプトを最適化することで、特定のドメインに適応させる手法である。モデルの重みを変更せずに適応が可能なため、コストが低く、迅速な実装が可能である。プロンプトの設計には、ドメインに関する深い知識や、モデルの動作特性を理解するスキルが求められる場合もある。この手法は他の手法と組み合わせることで、より効果的な結果が得られる。

（2）RAG（Retrieval-Augmented Generation）

RAG は、言語モデルと情報検索を組み合わせて応答を生成する手法である。

外部データベースから最新の情報を取得し、モデルの応答に組み込むことで、回答の精度と信頼性が向上する。実装にはベクトルデータベースの構築や検索アルゴリズムの最適化が必要だが、モデル自体の再学習が不要な点が利点である。ただし、検索精度や情報の整合性の管理が課題となる。

（3）教師ありファインチューニング

特定のデータセットでモデルのパラメータ（重み）を微調整し、特に出力形式の一貫性を重視する手法である。教師あり学習による微調整のため、応答品質の安定化に寄与する。特に、指示（Instruction）を理解し適切な応答を生成できるよう調整するインストラクションチューニングは、多種多様なタスクへの対応を可能にする。 モデルの重みを変更するので、一定の計算コストがかかる。

（4）継続事前学習

モデルに新たな言語やドメイン知識を追加学習させる手法である。専門知識をモデルに追加で学習させることが可能であるが、追加学習による計算コストが発生する。また学習の結果、Catastrophic Forgetting（過去の知識の喪失。新しい知識の追加学習により、過去の知識が失われる現象）や過学習のリスクがあるため、適切な管理が必要である。

（5）フルスクラッチ学習

フルスクラッチ学習は、特定のドメインに対してゼロからモデルを学習させる手法であり、最も高いコストと計算資源を必要とする。一方で、学習が成功した場合には、特定のドメインや業界に完全に最適化されたモデルを構築することができるため、高度なセキュリティや高い精度が求められる場合に有用である。

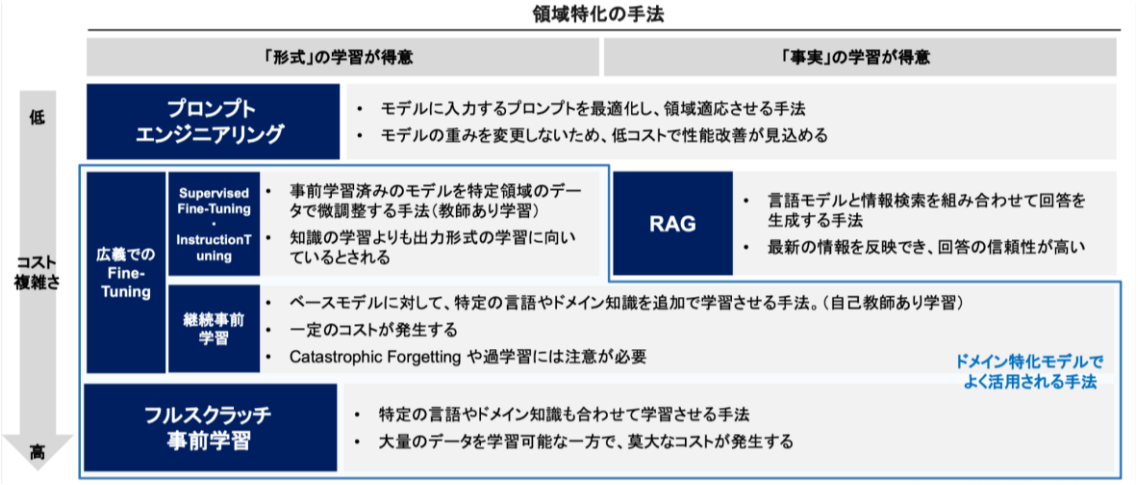


図 1　ドメイン特化の各種手法（筆者作成）

2.4 金融特化型 LLM の開発と応用

金融分野において、専門知識、金融の文脈理解を反映した LLM の開発が進んでいる。ドメインに特化させる主目的は、当該ドメインに関連したタスクの精度向上にあるが、これは金融分野において特に重要と言える。金融分野に固有の下記の特徴がその背景にある。

第一に、専門用語・知識が豊富に存在する点である。具体的には、金融業界特有の略語（IPO、M&A、ROE、EBITDA 等）や金融商品の専門用語、さらに Web 上の知識からは得られない実務的な知識・知見が多く存在する。

第二に、文脈依存性が高い点が挙げられる。例えば、「原油価格の上昇」という文書はエネルギー関連企業にとっては収益増加につながるためポジティブであるが、製造業などにとってはコスト増となりネガティブとなりうる。また「企業内で 1 万人規模の人員削減」という文書は、ネガティブな印象を持つ方も多いだろうが、投資家の観点からは短期的には人件費削減による財務改善につながるためポジティブに捉えられ、一方で長期的な成長力の低下を懸念してネガティブに解釈されることもある。このように、同一の事象であっても業界や時間軸によって評価が異なるため、金融ドメインに特化した深い文脈理解は重要である。

上記の背景をもとに、英語圏を中心として金融特化型 LLM の開発が進んできている。下記に幾つかの事例を掲載する。

- BloombergGPT：Bloomberg 社が開発した金融特化型 LLM。金融ニュース、レポート、市場データなど、幅広い金融データと一般的なデータをおおよそ半々にしてフルスクラッチで学習したモデルである。本モデルにプロンプトエンジニアリングを組み合わせることで、金融系のタスクも汎用的なタスクにおいても高い精度を示すことが示された。（2023 年前半に発表されたが、一般公開されていないのでその後の技術面の進化についての詳細は不明。）

- FinMA：Meta 社より公開された Llama という汎用 LLM を、金融データでインストラクションチューニングを実施した LLM。金融ニュースや企業レポートなどの金融関連のテキストから収集された文を対象に、センチメントを予測するタスクにおいて、GPT4 を上回る精度を達成。
- FinGPT：金融分野に特化したオープンソースの LLM であり、以下の特徴を持つ
  データ中心のアプローチ: インターネット上の公開データを活用し、特定企業の専有データに依存せず、広範な金融データを取り扱う。
  軽量な適応とコスト効率: 既存のオープンソース LLM（例：LLaMA、ChatGLM）を基盤とし、インストラクションチューニングを実施。
  多様な金融アプリケーションへの適用: ロボアドバイザー、運用取引、リスク管理、詐欺検出、信用スコアリングなど、幅広い金融タスクに応用する。

これらの金融特化型 LLM は、汎用 LLM と比較して、金融分野における特徴的な進化を遂げている。

まず、金融用語と概念の理解が向上し、特に金融関連文書のセンチメント分析において、従来の汎用 LLM と比べて予測精度が向上している。これは、金融ドメイン特有の専門用語や文脈の理解が深まったことによる。

次に、この理解力の向上を基盤として、実務応用の範囲が拡大している。例えば、大量の金融ニュースや市場データのリアルタイム解析によるトレンド把握や、企業の財務報告書と市場データの統合分析による投資判断支援など、より高度な業務への適用が進んでいる。さらに、分析の精度と包括性を高めるため、テキスト、数値データ、チャートなどの異なる形式のデータを統合的に処理するマルチモーダル化や、市場の時系列データの認識能力の向上が研究されている。これにより、より包括的で正確な金融分析が可能となることが期待される。

２.５ 金融特化型 LLM の今後の展望

金融特化型 LLM の開発は急速に進展しているものの、いくつかの課題も存在する。

- データの質と量：金融分野に特化した高品質なデータを十分な量で確保することは容易ではなく、モデルの性能向上において障壁となっている。
- モデルの更新：特に金融のような急速に変化する分野では、モデルを最新の情報で継続的に更新しなければ、その有用性が減少する可能性が高い。
- 説明可能性：金融のような重要な意思決定を伴う分野では、モデルの判断根拠を説明できることが重要である。

これらの課題に対する今後の展望として、以下の発展が期待される。

- マルチモーダル学習：テキストだけでなく、画像や音声なども含めた総合的な情報処理能力の向上。

- 継続的なモデルの更新 : 新しい情報を効率的に学習し、モデルを動的に更新する技術の発展。
- 説明可能性の向上 : モデルの判断プロセスをより透明化し、人間が理解しやすい形で説明する技術の進歩。

このような発展により、金融分野において、LLM はより高度で信頼性の高い支援ツールとしての役割を果たすことが期待される。

３．金融分野における LLM 活用の課題と対策

３.１ はじめに

LLM のビジネス活用を成功させるためには、多面的な評価により、LLM の信頼性の担保に努める必要がある。Huang.*et al*[21]では、LLM の信頼性を評価する観点として、8 つの主要な信頼性の側面（真実性、安全性、公平性、堅牢性、プライバシー、機械倫理、透明性、説明責任）について整理している。

金融分野における LLM の活用は、業務効率化や顧客サービスの向上に大きな可能性を秘めている。しかし、その特性と金融業界特有の要件から、実務での活用においては複数の課題が存在する。本章では FDUA による金融機関における生成 AI の実務ハンドブック[22]、FSB[23]に基づく文献調査と金融実務家へのインタビューから特に重要と考えた課題の「真実性、透明性、プライバシー」を取り上げ、その本質的な問題と対策について考察する。

３.２ 真実性（モデルの信頼性）

利用者保護や金融安定を目的に厳しい規制要件が課せられている金融分野での LLM 活用における最も根本的な課題は、モデルの信頼性である。LLM は確率的な生成モデルであり、時として「ハルシネーション」と呼ばれる誤った情報を生成してしまう現象を起こす。加えて、学習後に生じた新たな事象や最新の研究結果などはモデル内部に反映されず、また、学習データ自体に事実誤認や偏見・バイアスが含まれていれば、その情報を元にもっともらしいテキストが生成されてしまう可能性がある。金融取引や投資判断など、高い正確性が求められる領域では、この特性は自社の利益や顧客の信用を損なう重大なリスクとなる。

これらの課題に対する主な対策として、以下の方針が考えられる。第一に、LLM の精度的な限界を認識した上で、その役割を明確に定義することが重要である。金融取引などの重要な判断を伴う業務や、顧客へ直接的に接する業務においては、人間の専門家による適切な確認と修正を必須とする「ヒューマンインザループ」（人間が監視・介入可能な形での運用）体制を構築

---

21) Yue Huang., Lichao Sun., Haoran Wang., Siyuan Wu., Qihui Zhang., Yuan Li, *et al*. (2024) “TrustLLM: Trustworthiness in Large Language Models”, arXiv:2401.05561

22) 金融データ活用促進協会（FUDA） (2024) 金融機関における生成 AI の実務ハンドブック（第 1.0 版）

23) 金融安定理事会（FSB） (2024) The Financial Stability Implications of Artificial Intelligence https://www.fsb.org/uploads/P14112024.pdf

する必要がある。

第二に、モデル出力の精度向上のための技術的対策も重要である。具体的には、LLM に対して明確で具体的な指示（プロンプト）を与えることや、RAG を実施することで、ハルシネーションの発生確率を低減できる。例えば、社内の規程や過去の事例データなど、検証済みの情報を RAG で参照できるようにすることで、より正確な情報生成が可能となる。

第三に、複数の LLM を組み合わせたモニタリング体制の構築が有効である。主たる LLM の出力に対して、異なるアーキテクチャや学習データにより作られた別の LLM で検証を行うことで、出力の信頼性を高めることができる。例えば、LLM が生成した市場分析レポートの内容を、別の LLM が事実関係の確認や論理的整合性のチェックを行うといった運用が考えられる。このような重層的な検証体制により、単一の LLM による誤った情報生成のリスクを低減することが可能となる。

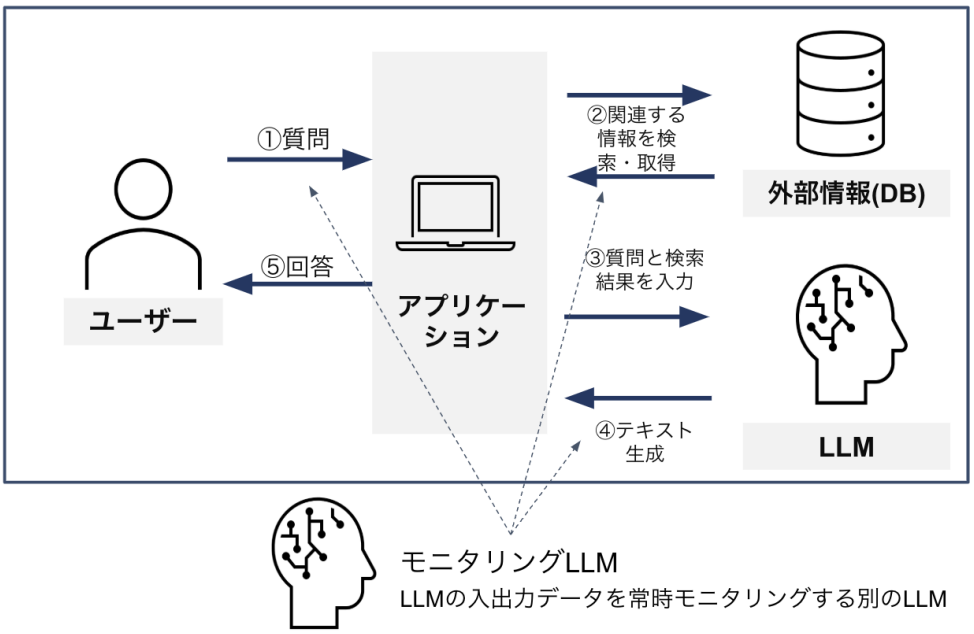


図 2　モニタリング LLM という考え方（筆者作成）

３.３透明性（ブラックボックス問題）

金融分野における LLM の活用において、モデルの出力のプロセスの不透明性、いわゆる「ブラックボックス問題」は非常に重要な課題である。金融機関の投資判断や与信判断などにおいて、その意思決定プロセスの説明責任が求められるが、LLM の出力結果がどのような根拠に基づいているのかを明確に示すことは一般に困難である。

このブラックボックス問題に対する対策として、主に二つのアプローチが有効である。第一に、プロンプトエンジニアリングの活用が挙げられる。LLM に対して、判断根拠や参照情報を明示的に出力するよう指示を組み込むことで、モデルの判断プロセスの透明性を高めることが

できる。例えば、「判断理由を 3 つ以上挙げ、それぞれの根拠となる具体的なデータを示してください」といった形で指示を与えることで、説明に寄与する出力を得ることができる。

第二に、段階的な判断プロセスの導入が重要である。具体的には、LLM の出力を最終的な意思決定として扱うのではなく、人間の専門家が判断を行う際の参考情報として位置づける。例えば、投資判断においては、LLM が提供する市場分析や企業評価を一つの情報源として扱い、他の定量的・定性的分析と組み合わせながら、最終的な判断は人間が行うというプロセスを確立する。これにより、判断の説明責任を明確に果たすことが可能となる。

３.４ プライバシー（セキュリティの課題）

金融機関はその性質上、個人情報（氏名、住所、電話番号、メールアドレスなど、個々の人物を識別できる情報）や機密情報（企業や組織の営業秘密、取引情報、内部戦略などの情報）などの秘匿性の高い情報を扱うため、LLM の利用におけるセキュリティ確保は重要課題の一つである。特に、クラウドベースの LLM サービスを利用する場合、データの越境移転や第三者アクセスのリスクが存在する。

加えて、LLM 特有のセキュリティ脆弱性として、プロンプトインジェクション攻撃とアドバーサリアル攻撃などが懸念されている。プロンプトインジェクション攻撃とは、悪意のある指示をプロンプトに含めることで、LLM の動作を操作する攻撃手法である。例えば、セキュリティ制限を回避して機密情報を漏洩させたり、不適切な出力を生成させたりする可能性がある。

一方、アドバーサリアル攻撃は、LLM の出力を意図的に操作するよう設計された入力を用いる攻撃手法であり、モデルの判断を誤らせることができる。これらの脆弱性は、外部の攻撃者による情報侵害や、AI の訓練データベースへの意図しないデータの混入などの問題を引き起こす可能性がある。

これらのリスクに対する対策として、まず組織的な管理体制の確立が重要である。具体的には、データモニタリングを行う専門チームの設置や、LLM に投入されるプロンプトの内容を厳密に管理するプロンプト監査体制の整備が必要である。また、技術的対策として、個人情報の匿名化技術の導入や、機密情報を自動的に検知・フィルタリングするシステムの実装が有効である。さらに、LLM の出力内容の機械と人によるレビューを組み合わせることで、セキュリティ対策の実効性を高めることができる。

組織的な対策としては、社員への定期的な教育・訓練プログラムの実施が重要である。このプログラムでは、生成 AI の適切な利用方法、機密情報の取り扱いルール、セキュリティリスクの認識向上などを包括的に取り扱う。また、入力データのリアルタイムモニタリングシステムを導入し、不正利用の早期検知と対応を可能にすることで、組織全体のセキュリティレベルを向上させることができる。

３.５ 結論

本章では、金融分野における LLM 活用の主要な課題として、モデルの信頼性、ブラックボ

ックス問題、セキュリティの三点について詳細な検討を行った。これらの課題は独立して存在するわけではなく、それぞれが相互作用する複合的な問題構造をなしており、金融分野特有の厳格な規制要件や顧客保護の必要性と相まって、包括的な対応策の策定が求められる。

これらの課題に対処するための重要なアプローチとして、RAG の活用により LLM の性能を引き出すことと、LLM の金融に関する能力の評価が挙げられる。

まず、RAG を活用することで、LLM が確率的生成モデルであるがゆえに直面する信頼性の課題、特にハルシネーション問題を効果的に軽減できる。RAG は外部の信頼性の高い情報源を参照することで、LLM の生成する出力の正確性を向上させる。この仕組みにより、金融分野で不可欠な精度と信頼性を確保できる。次に、ブラックボックス問題への対応も可能になる。外部情報源を明示的に利用するため、出力結果に対する説明可能性が高まり、LLM の判断根拠を可視化できる。さらに、信頼性の高い外部情報源を利用することで、誤情報や意図的な攻撃に対する耐性、セキュリティも強化できる。

同時に、LLM の金融に関する能力評価も重要である。LLM が金融ドメイン特有の専門用語や文脈の理解ができることは、正確性が求められる金融ユースケースにおいて欠かせない。さらに、金融面での性能評価ができれば、LLM と RAG の連携による出力の品質向上が客観的に測定でき、適切な改善指針を得ることが可能となる。したがって、LLM の性能評価と RAG の活用は、これらの課題を解決するための補完的かつ相乗的なアプローチとして位置づけられる。RAG による信頼性と説明可能性の向上、そして LLM 評価による改善指針の策定を組み合わせることで、金融業界における LLM 導入のハードルを下げると同時に、その運用価値を最大限に引き出すことが可能となる。

次章では、LLM の金融に関する性能を評価するための手法とその限界について詳しく検討し、さらに次の章以降で RAG の具体的な活用可能性を探る。

４．金融領域における LLM の評価手法と課題

４.１ はじめに

LLM の急速な発展に伴い、その性能を適切に評価することの重要性が増している。特に金融セクターでの LLM の利用にあたっては、モデルの誤作動や不適切な出力がビジネスや規制対応に重大な影響を及ぼす可能性があるため、適切なモデルの構築・利用と評価・検証、継続的なモニタリング等を通じたモデルリスク管理が不可欠である。ここでいうモデルリスク管理とは、モデルの誤り、不適切な使用、あるいは意図しない結果から生じる潜在的な損失を特定・評価・軽減するプロセスを指す。

本章では、まず LLM の一般的な評価手法について触れた上で、金融分野における LLM 評価の特有の課題と最新の評価手法について紹介する。

４.２ LLM の一般的な評価手法

４.２.１ 様々なベンチマークテスト

LLM の評価において、ベンチマークテストは最も広く用いられている手法である。これらのテストは、様々な自然言語処理タスクにおけるモデルの性能を定量的に測定するために設計されている。言語理解に関する主なベンチマークテストとして、以下が挙げられる。

（1）GLUE （General Language Understanding Evaluation）[24)]

GLUE は、自然言語理解に関連する 9 つの異なるタスクを含むベンチマークであり、文脈理解や推論能力を評価するために使用される。これには、自然言語推論（NLI）、文の類似性判断、感情分析などが含まれている。

（2）SuperGLUE[25)]

GLUE の後継として開発された SuperGLUE は、より難易度の高いタスクを追加し、LLM の限界をより深く探ることを目的としている。これには、常識推論や対話理解などのタスクが含まれ、モデルの言語理解能力をさらに高めるものである。

（3）LAMBADA[26)]

LAMBADA は、長文脈に基づいて正確な単語を予測できるかを評価するためのベンチマークである。これにより、LLM の文脈依存性の理解度が試される。LAMBADA は、文の最後に来るべき単語を予測することで、モデルの長文脈理解能力を測定するものである。

（4）SQuAD（Stanford Question Answering Dataset）[27]

SQuAD は、質問応答タスクに特化したベンチマークであり、与えられた文書から正確に質問に答えられるかを評価するものである。SQuAD データセットは、モデルの質問応答能力の評価に広く使用されており、後続のバージョンである SQuAD 2.0 では、回答が文書に存在しない場合も含めてモデルの性能を評価することが可能である。

そのほか、対話・生成系のベンチマークや、コモンセンス推論（一般的な知識や常識に基づいて行われる推論）のベンチマーク、マルチモーダルの評価など、LLM の進展に伴い LLM の評価方法も多様化してきている。さらに LLM 用の様々なタスクによるベンチマーク計測プラットフォームの Language Model Evaluation Harness なども登場し、ベンチマークの進展も著し

---

24) Alex Wang., Amanpreet Singh., Julian Michael., Felix Hill., Omer Levy. & Samuel R. Bowman. (2019) “GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING”, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* pp. 353-355

25) Alex Wang., Yada Pruksachatkun., Nikita Nangia., Amanpreet Singh., Julian Michael., Felix Hill., Omer Levy., Samuel R. Bowman. (2020) ”SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* Article No.: 294, pp. 3266-3280

26) Denis Paperno., German Kruszewski., Angeliki Lazaridou., Quan Ngoc Pham., Raffaella Bernardi., Sandro Pezzelle., Marco Baroni., Gemma Boleda., Raquel Fernandez. (2016)“The LAMBADA dataset: Word prediction requiring a broad discourse context” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Volume 1: Long Papers, pp. 1525-1534

27) Pranav Rajpurkar., Robin Jia., Percy Liang. (2018) “Know What You Don’t Know: Unanswerable Questions for SQuAD” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* Volume 2: Short Papers, pp. 784-789

い。これらのベンチマークテストは、LLM の言語理解能力や推論能力を多角的に評価することを目的としており、LLM の評価における重要な手法と言えるが、一方で評価するためのデータセットの作成には、場合によっては人間による評価の過程が必要であるなど、一定の労力を要する。

４.２.２ LLM-as-a-Judge

そこで出てきたのが、LLM-as-a-Judge[28)]という、LLM の出力を、同じく LLM を用いて評価する手法である。従来、LLM の出力評価は人手やルールベースで行われてきたが、LLM の出力は確率的で多様性が高いため、これらの方法には限界があった。LLM-as-a-Judge は、LLM 自身の高度な言語理解能力を活用し、出力の妥当性や品質を効率的かつ一貫して評価することを可能にする。

この手法では、評価基準を明確に定義し、それをプロンプトとして LLM に与えることで、出力の正確性、簡潔さ、適切な言葉遣いなど、サービス要件に沿った多角的な評価を行うことができる。特に、GPT-3.5 や GPT-4 などの高度なモデルの登場により、LLM-as-a-Judge の有用性はさらに高まっている。

LLM-as-a-Judge の主な利点は以下の通りである。

- 効率性：人手による評価と比較して、迅速かつ大量の出力評価が可能である。
- 一貫性：評価基準を統一することで、評価結果のばらつきを抑えることができる。
- 柔軟性：評価基準の変更や追加が容易であり、様々なタスクや要件に対応できる。

LLM-as-a-Judge を導入する際には、評価基準の明確化やプロンプト設計の工夫が重要となる。また、LLM 自身のバイアスや限界を考慮し、評価結果の解釈には注意を払う必要がある。LLM-as-a-Judge は、LLM の出力評価において効率的かつ柔軟な手法として、今後の LLM アプリケーション開発や運用において重要な役割を果たすことが期待される。

４.２.３ 金融分野の LLM の評価について（文献調査）

金融のような特定の分野での活用を前提とすると、LLM の評価には、特定のタスクや応用分野に焦点を当てた手法が不可欠である。これらの評価手法は、各分野の固有のニーズに応じた指標を用いて、LLM の実用性と効果を測定するために開発されている。タスク特化型評価手法は、各分野の専門的なニーズに応じた指標を用いることで、LLM がその分野においてどの程度有用であるかを適切に評価することを可能にする。金融、医療、法務、科学技術といった異なるドメインでは、それぞれ特化型のモデルが用いられ、その分野固有のタスクに応じた評価が実施されている。FinBERT[29)]の研究においては、金融関連のニュース文をポジティブ、ネガティ

---

28) Lianmin Zheng., Wei L. Chiang., Ying Sheng., Siyuan Zhuang., Zhanghao Wu., Yonghao Zhuang., Zi Lin., Zhuohan Li., Dacheng Li., Eric P. Xing., Hao Zhang., Joseph E. Gonzalez., Ion Stoica. (2023) “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”, arXiv:2306.05685

29) Dogu Araci. (2019) “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models”, arXiv:1908.10063

ブ、ニュートラルの三値に分類するタスクと、金融ニュースヘッドラインやX（旧 Twitter）のツイートのセンチメントをスコアとして（-1 から 1 の連続値）を予測するタスクが金融 LLM の主要な評価として用いられていた。

その後、金融分野における LLM の評価についても LLM の進展と合わせて高度化が進み、Shah *et al.*[30]は、FLUE（Financial Language Understanding Evaluation）という、金融分野に特化した自然言語処理タスクを評価するための包括的なベンチマークが提唱しており、以下の 5 つのタスクが含まれている。

- 金融感情分析（Financial Sentiment Analysis）（分類・回帰あり）
- ニュースヘッドライン分類（News Headline Classification）
- 固有表現抽出（Named Entity Recognition, NER）
- 構造境界検出（Structure Boundary Detection, SBD）
- 質問応答（Question Answering, QA）

	説明	評価指標
FPB	• FPB Sentiment Classification（Financial Phrase Bank感情分類） • FPBデータセットを用いた感情分類タスク。金融ニュースなどの文書内のフレーズが、ポジティブ、ネガティブ、ニュートラルなどの感情カテゴリーに分類	Accuracy
FiQA	• FiQA Sentiment Analysis（FiQA感情回帰分析） • FiQAデータセットに基づく感情回帰タスク。このタスクでは、金融関連のテキストから感情スコアを予測し、実際のスコアとの誤差を測定	MSE
Headline	• Headline Classification（ニュース見出し分類） • ニュース見出しを複数のカテゴリに分類するタスク。金融ニュースには株価の上下など、株価に関連するカテゴリが含まれる	F1
NER	• NER（Named Entity Recognition、固有表現認識） • 金融テキスト内の固有名詞を分類するタスクで、会社名、人名、地名などのエンティティを識別	F1
FinSBD3	• FinSBD3（Structure Boundary Detection、構造境界検出） • 金融文書の構造（見出しやリスト項目の境界）を検出するタスク	F1
FiQA QA	• FiQA QA（質問応答） • 金融関連の質問に対する適切な回答を予測するタスク	nDCG（Normalized Discounted Cumulative Gain）

図 3　金融分野での LLM 評価用データセットと評価指標（筆者作成）

さらに近年、金融特化型 LLM の開発が進展する中で、より多様な金融アプリケーションへの適用が模索されており、その性能を評価するためのベンチマーク研究が登場し始めている。

特に、FinBen[31]はその代表例として挙げられ、金融分野における LLM の能力を包括的に評価するために設計されたベンチマークであり、情報抽出、テキスト分析、質問応答、テキスト生成、リスク管理、予測、意思決定など 7 つの主要なタスク領域をカバーしている。

---

[30] Raj Sanjay Shah., Kunal Chawla. , Dheeraj Eidnani., Agam Shah., Wendi Du., Sudheer Chava., Natraj Raman. , Charese Smiley., Jiaao Chen., Diyi Yang. (2022) “WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain”, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* pp. 2322-2335

[31] Qianqian Xie., Weiguang Han., Zhengyu Chen., Ruoyu Xiang., Xiao Zhang., Yueru He. *et al*. (2024) “FinBen: A Holistic Financial Benchmark for Large Language Models”, arXiv:2402.12659

		評価指標	検証タスク
評価・検証方法	情報抽出（IE）	・金融文書内の重要なエンティティや関係を特定し、非構造化データを構造化されたインサイトに変換可能か評価	・金融契約書やSEC申請書からエンティティを抽出。因果関係の分類や数値ラベル付けなども実施
	テキスト分析（TA）	・金融テキストの内容や感情分析を行い、市場動向の理解の補助となるか評価	・金融テキストから感情や意見を抽出し、価格行動の分析や論点分類などを実施
	質問応答（QA）	・金融関連のクエリを理解し、応答する能力を評価	・複雑な金融クエリに対し数値推論や複数回の対話形式の質問を実施
	テキスト生成（TG）	・一貫した金融テキストを生成する能力を評価	・情報量の多い金融テキストの生成を実施
	リスク管理（RM）	・信用リスクの評価、不正行為の検出、規制遵守の確保などを網羅的に評価	・信用スコアリング、不正検出、財務破綻予測などのリスクに関連する情報の識別・抽出・分析を実施
	予測（FO）	・将来の金融動向を予測し、市場のダイナミクスに対して戦略的な対応が可能か評価	・株価の動きや市場の動向の予測を実施
	意思決定（DM）	・取引戦略の策定や投資ポートフォリオの最適化など、情報に基づいた金融意思決定を行う能力を評価	・取引戦略の策定、ポートフォリオ最適化などの意思決定を実施

図 4　金融領域での LLM 評価項目（筆者作成）

		定義	評価対象指標
評価方法	F1スコア	・PrecisionとRecallの調和平均	・情報抽出、テキスト分析、QA、リスク管理
	Accuracy	・全ての予測結果のうち、正しく予測されたものの割合	・テキスト分析、予測
	RMSE	・予測値と実際の値の間の平均の誤差	・テキスト分析
	AveF1	・異なるクラスやタスクにおけるF1スコアの平均	・テキスト分析
	EntityF1	・固有表現認識におけるF1スコアの総和	・情報抽出
	EmAcc	・モデルの予測が正解データと完全に一致する割合	・情報抽出、QA
	ROUGE	・生成されたテキストと参照テキストの類似性	・テキスト生成
	BERTScore	・BARTモデルを使用し生成されたものと参照テキストの類似度	・テキスト生成
	MCC	・2値分類における正確さをTP,TN,FP,FNで評価する指標	・リスク管理、予測
	SR	・投資においてリスクを取って得られるリターンの割合	・意思決定

図 5　評価の方法と用いられる評価指標（筆者作成）

注目すべきは、従来のベンチマークが対象外としてきた株価予測や金融取引における意思決定といった高度な金融タスクを評価範囲に含めることで、LLM が金融特有のシナリオに対してどの程度適応可能であるかを包括的に測定する枠組みを提供している点である。また、評価指標に関しては、従来の機械学習や統計的手法に基づく指標に加え、取引戦略（意思決定）については、Sharpe Ratio（シャープレシオ）といった金融分野における専門的な評価指標を採用することで、モデルのパフォーマンスが実際の投資環境においてどの程度有用であるかを実践的に評価する新しいアプローチを示している点も特徴的である。

LLM の実務的なタスク適応力をより精緻に測定できるため、今後の金融特化型 LLM の研究・開発をする際の評価の基盤としての利用が予想され、実際に、金融分野の LLM の性能を評価・

比較するプラットフォームである Open Financial LLM Leaderboard[32]では、FinBen で提唱された 7 つの主要なタスク領域について多数の LLM の評価が行われている。

Model	Average	Average IE	Average TA	Average QA	Average TG	Average RM	Average FO	Average DM	Sauce
GPT4-turbo	39.2	35	64.4	50.7	10	51.7	54.3	75.2	Close
LLaMA3.1-70B	36.2	15.7	63.6	14.7	9	0	46	49.3	Open
Qwen2-72B	34.7	12.6	59.5	0.3	11	0	53.7	0	Open
Xuanyuan-70B	34.4	9.3	61.4	0.7	12.5	0	51.7	0	Open
LLaMA3.1-8B	34.3	15.6	56.2	1.3	10	0	54.3	0	Open
Gemini	32.4	22.1	58.4	20.3	19.5	51.8	53.7	67.2	Close
ChatGPT	29.2	26.4	59	39.3	8.5	45.6	52.7	0	Close
meta-llama/Llama-2-70b	25.8	10.6	59.9	10.7	12.5	50	49	0	Open
Duxiaoman-DI/XuanYuan-6B-Chat	25.7	11.1	54.2	3.7	12	50.7	50.3	0	Open
Qwen/Qwen2-7B-Instruct	22.9	9.9	52.7	0	11	51.6	52.3	0	Open
TheFinAI/finma-7b-full	21.5	12.6	48.7	8	6.5	49.7	50.7	0	Open
internlm/internlm-7b	20.4	12.6	47.3	0	6.5	50.2	54.7	0	Open

図 6　Open Financial LLM Leaderboard の上位 12 個の結果（筆者作成）

結果については、GPT4-turbo が情報抽出タスクを中心に、全体的な強さを見せる一方で、Gemini もテキスト生成やリスク管理といった高度な推論を伴うタスクが強いことが伺え、汎用 LLM が金融特化型 LLM に比べて、以前優位性があることが確認できる。また、日本における研究として、平野[33]は、日本語の金融分野に特化した LLM ベンチマークを構築し、GPT-4 など主要なモデルの性能を評価している。感情分析や証券分析、公認会計士試験、ファイナンシャルプランナー試験、証券外務員試験の模擬問題といった日本語の金融分野における 5 つのタスクが設定されており、日本市場における LLM の能力を測定する重要なツールとして機能しうる。

32) Huggingface (2024) Open Financial LLM Leaderboard https://huggingface.co/spaces/finosfoundation/Open-Financial-LLM-Leaderboard

33) 平野 正徳 (2023)「金融分野における言語モデル性能評価のための 日本語金融ベンチマーク構築」, jxiv.564

		タスク概要	検証方法
評価・検証方法	chabsa	• 金融文書の一種である、有価証券報告書に含まれる文章に関して、特定の単語に対するセンチメントを判定するタスク	• センチメントの分類として、PositiveとNegativeの二値分類を取り評価値としてそれぞれのmacro-f1値で評価
	Cma basics	• 証券アナリスト試験のサンプル問題をクローリングにより取得し成型したデータセットで構築された証券分析における基礎知識を問うタスク	• 証券アナリスト試験から図を含む問題を削除し選択形式で回答を問わせる方式で正答率を評価
	Cpa audit	• 公認会計士試験における短答式試験監査論の問題を収録したタスク	• 6択の問題を360問、5択の問題を38問取得しマーク式で回答をさせ正答率を評価
	fp2	• ファイナンシャルプランナー試験2級の選択問題を回答させるタスク	• 2021年5月から2023年9月の過去問題を公式HPより取得し、図の問題の削除,表はマークダウン形式と問題を成型した問題への回答率を評価
	Security sales 1	• 証券外務員試験1級に相当する模擬試験のタスク	• 外務員試験1級の文字試験や対策問題例をクローリングし、図の問題の削除などの成型をした問題への回答率を評価

図 7　日本語を対象とした金融領域での LLM 性能評価項目（筆者作成）

４.２.４ 金融分野の LLM の評価について（インタビュー調査）

筆者らが行った AI 研究者、AI 開発者、金融実務家へのインタビュー調査の結果によると、日本の金融機関での LLM 適用の現場においては、汎用 LLM を用いる事例が支配的であった。そして、ユースケースごとに、実務経験豊富なメンバーが QA セット（質問・回答セット）を作成し、LLM による出力と事前に用意した模範解答を比較する方法が多く採用されている。この QA セットは、各業務における典型的なケースから特殊なケースまでを網羅するように作成することが多い。

この評価方法では、LLM の出力が模範解答とどの程度一致するかを指標化し、モデルの実務対応能力を客観的に評価している。こうした定量的な評価手法により、LLM が特定の業務においてどの程度役立つのかを客観的かつ精緻に測定することが可能となる。このような評価結果を経営層や上長に対して明確に説明できることから、金融機関での実務適用における信頼性の確保に大きく寄与していると思われる。

一方で、このような評価手法には、QA セット作成に要する工数の問題が付きまとう。例えば、QA セット作成の効率化のために、過去の業務記録や相談事例などの既存データを活用し、半自動的に QA セットを生成する手法の開発も有効であると考える。また、複数の金融機関で共通して利用可能な基本的な QA セットを業界団体等で共同開発し、各社がそれを自社の特性に応じてカスタマイズするというアプローチも検討に値する。

４.３ 金融分野における LLM 評価の課題

文献調査とインタビュー調査を通じて、金融分野における LLM 評価の現状と課題について包括的な理解を得ることができた。一方で、未だ解決されていない課題も浮き彫りとなった。そこで本節では、金融領域の特性を踏まえ、現状の評価手法の課題と今後の展望について検討する。

最も基本的な課題として、データの絶対量の不足が挙げられる。オープンソースの金融デー

タに共通する問題として、利用可能なデータセットの規模が小さい点は、様々な文脈におけるモデルの評価を行う上で大きな制約となっている。この課題に対しては、より多種多様なデータセットの開発・確保が必要不可欠である。これらのオープンソースデータをもとに、より広範なタスクも作成が可能である。

言語対応の観点からは、日本語での評価基準の確立が重要な課題となっている。既存の研究の多くは英語での評価が主流であり、日本語での金融特化型の評価指標は極めて限定的である。国内企業を中心に、LLM の日本語評価の基準の整備は進みつつあるが、金融領域に特化したものとなると、データの量的制約も加わり進展が十分でない。

技術的には、動的な市場環境への対応も重要な課題である。金融市場は常に変動し、ボラティリティの急激な変動や経済的ショック、規制変更に迅速に対応できるモデルとその評価が求められる。従来の評価手法の多くが静的なデータセットに依存しているため、市場環境の変化に応じた LLM の適応能力を測定することが困難である。現状では、モデルのリアルタイムデータに基づく判断能力の評価が不十分となり、実世界での適用性が正確に反映されないという課題が依然として存在する。

４.４ 結論

LLM の評価手法は急速に発展しているものの、特に金融分野における適切な評価方法の確立には多くの課題が残されている。特に日本語においては評価手法が乏しく、今後の発展のためには、金融専門家と AI 研究者の協力が不可欠であり、業界標準の評価フレームワークの開発や、実際の金融環境での長期的な性能検証が必要となるだろう。今後、これらの課題に取り組むことで、金融分野における LLM の信頼性と有用性が向上し、より安全で効果的な活用が可能になると期待される。

５. RAG（Retrieval-Augmented Generation）の概要と性能

３、４章で述べた LLM が持つ性質の一つである、事実にもとづかない内容を生成してしまうハルシネーションの問題への対応、AI 利用の信頼性を高めるためのアプローチとして、情報検索と LLM を組み合わせた RAG（Retrieval-Augmented Generation）という技術を用いることが考えられる。RAG を用いたシステムは、ユーザーからのクエリに対して関連する社内の文書などクローズドなデータ（マニュアルなどの非構造データ）を検索し、その情報を基に正しい応答を生成することを可能にする。これにより、検索の高度化による情報収集の効率化や業務負担の軽減、顧客満足度の向上といった効果が期待される。本節では、RAG 技術の基本的な概念から、金融機関における実用性、導入における課題、及び応用例について考察し、今後の可能性を検討する。

５．１ RAG の仕組みと特徴

RAG という言葉は、Lewis *et al.*[34]により初めて提案されたものである。RAG とは、クエリに対する回答を生成する際に、クエリに関連するデータを検索し、その情報を回答生成用のプロンプトに追加して最終的な回答を生成する手法である。Gao *et al.*[35]による RAG の説明によると、図 8 にあるように、事前に、Indexing により整理された文書のデータベースから、ユーザーのクエリと関連する文書を検索（Retrieval）し、検索結果を回答の生成のためのプロンプトに追加、このプロンプトを LLM に入力することで最終的な回答を得るというものである。このように外部情報をプロンプトに加えた上で LLM の回答を生成させることにより誤情報の生成を防ぐ、適切な情報を回答させる、という用途に活用される。OpenAI 社による ChatGPT のような LLM は、事前に大規模なウェブ上のテキストデータをもとに学習をしており、対話や要約、翻訳などさまざまな言語に関するタスクを実施することができる。一方、事前に学習していない内容については正しい回答をすることは通常できない。LLM は事前に学習していない内容について、前に続く文章の内容に従って、もっともらしい内容を生成するが真実とは限らない内容を生成してしまう、ハルシネーションを起こしてしまうこともある。社内の業務に活用する場合には社内の情報を学習していない LLM に対して追加でデータを与える、学習をさせるなどの工夫を行い、ハルシネーションを起こさないようにする対策、組織内独自のデータに関連する応答も可能にさせるための対応が必要になる。各社独自での LLM 導入の考え方としてプロンプトエンジニアリングによって LLM の出力をカスタマイズする方法から、RAG を用いて社内のデータを活用したシステムを開発する方法、データを用いて、LLM を独自にファインチューニングにより開発するという方法などがある。ただし、独自の LLM の開発を行う場合は大量のデータの用意及び学習のコストが必要になる。金融機関の業務においては規制要件の変更等に伴い社内マニュアルの変更も比較的頻繁に行われることが考えられるため、その度に LLM を学習し直すのは現実的ではない。

このような課題に対し、RAG を用いることで大きなコストをかけることなく、社内用データを参照したシステムを開発できる。マニュアルが変更された場合にも、検索用のデータベースの変更で対応が可能であるという利点があり、RAG の開発を行う取り組みが増えてきている。

社内文書を対象とした RAG を用いたシステムを開発するという取り組みにおいて重要な技術モジュールは、データ整理（Indexing）、データ検索（Retrieval）、回答生成（Generation）の 3 つに分けて考えることができる。

（１）データ整理（Indexing）

社内文書、データは通常、PDF や Word ファイルなど様々な形式の図表やテキストを含むフォーマットのデータとして存在している。これらを RAG のシステムとして検索しやすい形に

---

[34] Patrick Lewis., Ethan Perez., Aleksandra Piktus., Fabio Petroni., Vladimir Karpukhin., Naman Goyal.*et al*. (2020) "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems,* Article No.: 793, pp. 9459-9474

[35] Yunfan Gao, Yun Xiong., Xinyu Gao., Kangxiang Jia., Jinliu Pan., Yuxi Bi. *et al*. (2023) "Retrieval-Augmented Generation for Large Language Models: A Survey", arXiv:2312.10997

整形するという処理を行う必要がある。一つの方法として、これらの非構造化データを一定の単位で分割（チャンク化）し、埋め込み（embedding）モデルを用いてベクトルデータに変換、ベクトルデータベースへ保存するという処理を行う。データ整理においては、テキストデータのチャンク方法と、利用する埋め込みモデルの選択という二つの工夫のポイントがある。テキストのチャンク方法に関して、一定の文字数ごとや、句読点をベースとした文章単位、章や節などパラグラフの単位で分割するという単純な方法から、文章の意味合いを考慮して論理的連続性を考慮した分割を行う方法などが存在する。通常、後者のような文章の意味を考慮して分割する方がユーザーの体感として精度が良いということが多いが、意味合いを考慮し文章の分割を人間が行う作業には大きな労力を要するため、意味合いを考慮して分割するということ自体も LLM に実施させるという取り組みも考えられる。埋め込みモデルの選択に関して OpenAI 社の text-embedding-3[36]や Wang. *et al.*[37]による多言語対応したオープンソースのモデルの Multilingual E5 などが存在する。

（2）データ検索 （Retrieval）

データ検索では、クエリに関連する文書を特定して取り出す役割を担う。検索の精度と速度はシステム全体の応答品質に直結するため、自然言語処理（NLP）やベクトル検索技術を用いて、クエリの意図を正確に理解し、最適な情報を取り出すことが求められる。ベクトル検索を用いる場合、1 で利用した埋め込みモデルと同じモデルを用いてクエリの内容をベクトル化し、ベクトルデータベースを検索、類似性が高い上位のチャンクを取得するというアプローチをとる。また、文書検索のための一連の検索システム（検索パイプラインと呼ぶ）には事前にフィルタリングやスコアリング機能を組み込み、無関係な文書の混入を防ぎ、効率的に最も有用な情報を取得するということも可能である。検索により得られた文書はコンテキストと呼ばれ、回答生成用プロンプトに何らかの形で加えられる。

（3）回答生成（Generation）

回答生成用プロンプトには、検索パイプラインで得られたコンテキストが組み込まれ、具体的でわかりやすい応答を LLM から得られるように構築される。LLM は、プロンプトに与えられた情報を基に自然な言葉での応答を生成する。回答の方法はユースケースに応じて様々な設計が必要になるが、いずれの場合もプロンプトの調整により柔軟な対応が可能である。コンテキストに含まれる情報のみを用いて回答させるように制限することや、LLM が事前に学習している知識も組み合わせて回答をするように指示を与えるということも可能である。ハルシネーションを抑えるためには、通常前者のような、コンテキストの内容のみを用いて回答するようなプロンプトを与えることが多い。

---

[36] OpenAI (2024) New embedding models and API updates https://openai.com/index/new-embedding-models-and-api-updates

[37] Liang Wang., Nan Yang., Xiaolong Huang., Linjun Yang., Rangan Majumder., Furu Wei. (2024) “Multilingual E5 Text Embeddings: A Technical Report”, arXiv:2402.05672

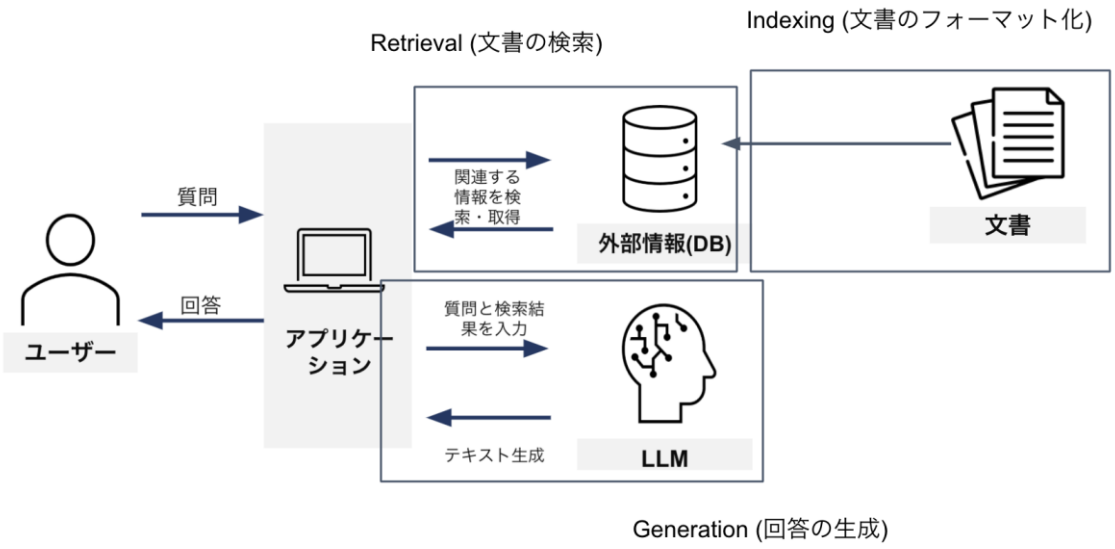


図 8　RAG の 3 つのコンポーネントの関係（筆者作成)

以上、Indexing、Retrieval、Generation の 3 つのモジュールにより RAG は構成されており、それぞれの部分においてユースケースに合わせて十分な精度を達成するための工夫が必要となる。

５.２ RAG の性能評価指標と測定方法

5.1 節で述べたように、RAG は、データ整理（Indexing）、データ検索（Retrieval）、回答生成（Generation）という３つのコンポーネントからなる。このうち、Retrieval における検索性能、生成された回答内容の品質を評価する方法に関するリサーチとして Yu *et al.*[38]のフレームワーク（A Unified Evaluation Process of RAG (Auepora)）を紹介する。

5.2.1 A Unified Evaluation Process of RAG （Auepora）による RAG 評価の手法について

本手法では、RAG のシステムのうち性能評価の対象として Retrieval 部分における検索の性能、回答の内容の 2 つを主な評価対象としている。具体的には入力データと正解データのペアとして、1. Retrieval による検索結果の文書と想定の回答に必要な文書候補、2. RAG システムによる回答内容とサンプルの正解となる回答、3. RAG システムによる回答に対し、後処理をした出力と正解のラベルという 3 つのデータの組み合わせを用いて性能評価を行うことを提案している。これらのデータのペアに対して、以下の精度指標を用いてシステムの性能の評価を行う。

（1）データ検索(Retrieval) の精度評価指標

[38] Hao Yu., Aoran Gan., Kai Zhang., Shiwei Tong., Qi Liu., Zhaofeng Liu. (2024) “Evaluation of Retrieval-Augmented Generation: A Survey”, arXiv:2405.07437

① 関連文書とクエリの関連性（Relevance）：検索により得られた文書がクエリの内容のために必要な情報とどれだけ一致するかを評価する指標

② 関連文書と想定の回答に必要な文書との関連性（Accuracy）：検索により得られた文書が必要な回答のための想定の文書の候補とどれだけ合致しているかを評価する指標

（2）回答生成（Generation）の精度評価指標

① 応答の内容とクエリの関連性（Relevance）：ユーザーからのクエリに対して、システムの応答がどれだけ回答として相応しいかを評価する指標

② 応答の内容と、参照ドキュメントの関連性（Faithfullness）：生成された応答が検索により得られた関連文書の情報を正確に反映しているか、整合性を評価する指標

③ 応答の内容と、想定の回答との関連性（Correctness）：生成された応答がサンプルの正解とする応答の内容とどれだけ類似するかを評価する指標

これら 5 つの指標を用いてシステムとしての正確性、有用性を評価するというフレームワークとなっている。また、この他に実運用を想定した他の測定対象の指標として Latency（システムの応答までにかかる時間、ユーザーにとって重要な指標となる）や Diversity（様々な関連文書を見つけたような応答が可能であるかどうか）、Noise Robustness（システムにとって関連しない文書の影響をどれだけ除外できるか）、Negative Rejection（利用可能な情報が不十分な場合に回答を行わないようにする）といった項目による、システム自体の有用性を評価する方法も挙げられている。

次に、金融領域の文書を対象とした RAG の性能評価の研究として、Zhao *et al.*[39]による調査を取り上げる。

5.2.2 Zhao *et al.*による金融領域の調査について

この調査では、2 つの実世界の金融領域のデータセットを用いて 15 の検索シナリオにおける RAG の生成内容の評価を行なった。実験では、次の 2 種類のコーパスからクエリを抽出し、RAG データセットを作成した。(1) 銀行のウェブページ：銀行商品に関する一般的な情報を提供する公開ウェブページ。(2) 銀行の業務方針ガイド：顧客サービス担当者向けの内部ガイドで、顧客支援に関する方針や手続きの詳細を含む。これら 2 つのコーパスに関連する質問は、実際のユーザークエリを基に、生産システムのログから収集したもの、または専門家が生成したものである。各ウェブページや記事は約 100 語単位のチャンクに分割し、文章のまとまりを保った状態で「ドキュメント」として扱った。特に検索条件に関する評価の内容について取り上げる。RAG システムにおいて、高い性能の回答を得るためには、回答のために必要な文書を

[39] Yiyun Zhao., Hanoz Bhathena., Prateek Singh., Saket Sharma., Bernardo Ramos., Aviral Joshi., Swaroop Gadiyaram. (2024) “Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 6: Industry Track, pp. 279-294

検索により正しく取得することが重要である。検索品質の違いによる回答内容の違いについて、本ペーパーでは以下のシナリオでテストを実施している。

（1）回答のために必要なドキュメント（ゴールドドキュメント）が取得されない場合の影響

Retriever-Only（ゴールドドキュメントが含まれていない可能性があるセット）と Retriever-W-GT（ゴールドドキュメントが確実に含まれるセット）の 2 つの条件を作成し、ゴールドドキュメントの有無が回答生成に与える影響の評価を実施している。

（2）取得ドキュメント数の影響

クエリに対する検索結果として条件に一致する上位数個のドキュメントを取得、コンテキストとして回答生成用プロンプトに追加する。ここでは、取得するドキュメント数を変えて LLM の回答生成を評価するため、top_3、top_5、top_20 の 3 つの条件を設定し評価を実施している。

（3）ドキュメント表示順序の影響

取得したドキュメントをプロンプトに追加する際の、順序に関する評価を実施している。Retriever-Only のケースでは、ドキュメントの順序をランダムにシャッフルした条件を追加し、順序の影響を評価している。Retriever-W-GT では、ゴールドドキュメントを最初または最後に配置した条件も追加し、順序に対する LLM の応答の敏感性の評価を実施。

以上のケースをまとめて、15 の検索条件を設定した。そして、RAG システムの性能を評価するために、以下の 2 つの側面から評価を行った。

- 回答の質 : タスクの内容は質問応答であるため、生成された回答と正解の回答との間で、一致する単語の精度を確認する指標を用いて実施している。
- 指示の遵守 : RAG では、LLM に特定の言語スタイルや構造化された形式での回答を要求することがある。例えば、引用の形式や回答のフォーマットなどである。したがって、モデルがこれらの指示をどの程度遵守しているかを評価している。

RAG システムにおける回答の質及び指示の遵守度に対するゴールドドキュメントの影響と、取得ドキュメント数の影響、ドキュメントの表示順の影響として次のような内容が明らかになった。

（1）ゴールドドキュメントが取得されない場合の影響について

回答を含む検証済みの文書（ゴールドドキュメント）が検索結果に含まれると、すべてのモデルで性能が向上することが明らかになった。一方、LLM の限界として、GPT-4 のような強力な LLM でも、検索失敗時に「回答が見つかりません」と適切に応答することは不完全であった。LLM ベースの RAG システムには、関連情報が文脈に欠けている場合でも、何らかの回答を生成してしまう、RAG システムに期待される、検索結果に含まれる文脈内の関連情報のみを使用して回答するということに反した回答を生成してしまう、という傾向が認められた。この傾向は結果として、質問に直接答えていない関連コンテンツを含む回答を生成してしまう、ハルシネーション（根拠のない回答の生成）を起こすというような問題を生じさせる。これらの結果は、RAG システムの設計において、正解文書の存在が重要であることと、過剰な回答を生

成してしまうという問題をどのように制御するかが課題であることを示唆している。

（2）取得ドキュメント数の影響について

検索結果として、RAG の回答生成プロンプトに加えられる文書について、もし、正解文書が含まれている場合には、追加される文書の数が増えるほど性能は低下する傾向が認められた。つまり、正解の文書が存在する場合に、その他の関連しない文書が混ざることは好ましくない、ということが明らかになった。逆に、正解の文書が含まれない場合、回答生成プロンプトに対して、文書数が増えるほど性能が向上するという傾向が得られている。2 点目に関する理由の分析として、多数の文書がコンテキストとして与えられた場合、正解に関連する内容がなんらかの形で含まれる可能性が高くなる可能性が高まるということ、及び、この文書セットの中に正解情報がある場合 LLM はそれを見つけることが得意であるという可能性が示唆されている。つまり、正しい情報が明らかに含まれるのであればその他の余分な情報は少ない方の性能が高いということがわかった一方、もし明確な正解が明らかでない場合は多くの情報を与えることで中から適切な情報を抽出、回答させる方の精度が良いということが明らかになった。

（3）検索文書の順序の影響

検索結果より得られたドキュメントを与える順番に関する分析として、順序は LLM の性能に顕著な影響を与えるという結果が得られている。正解文書を最初に配置すると、最後に配置するよりも性能が向上しており、特に検索文書数が多いほど、この差は顕著になる。

この調査の結論として、以下の点が明らかになった。

- 回答生成プロンプトに対して、多くの文書を追加するということでは精度向上の取り組みとしては不十分であり、LLM はノイズ文書に敏感であり、検索の最適化には他の要素も考慮が必要ということが明らかになった。
- ドキュメントを追加する順序及び、ノイズを除くことの重要性が明らかになった。関連して、検索結果に対する再ランキングシステムの重要性について、LLM ベースのパイプラインでも、検索結果の再ランキングは必要であり、特に文書数が多い場合に重要であるということが明らかになった。

これらの結果は、RAG システムの設計において、単に関連文書を検索するだけでなく、その数と順序を適切に制御することの重要性を示唆している。このような RAG システムの評価方法及び、性能の改善のための取り組みについて実際の金融機関での活用を想定し、金融庁関連の法令や監督指針を用いてデータセットを構築し検証を行う。

６. LLM と RAG の性能向上に向けた取り組み

RAG を用いた社内用の対話システム開発について多くの企業が取り組みを進めていることは 4.2 節でも述べたが、一方で思うような回答が得られない、回答の精度が不十分であるといったことが原因で実導入まで至らないというケースも多く存在している。特に金融機関の場合では特有の専門用語が多く含まれる、既存のドキュメント量が膨大に存在している、ドキュメ

ントの内容の変更が頻繁に行われるなどの原因により検索フェーズがうまくいかない、回答のために必要な文書をうまく抽出できないということが言われている。このような課題に対してどのような対応が考えられるか整理していく。Zhao *et al.* の調査では、金融領域の文書を対象とした RAG システムの性能評価に関して、検索フェーズ、回答生成フェーズなどの性能への影響を分析している。

本ペーパーでは、金融庁関連の過去の対応記録や法令・ガイドライン等の文書データを用いてテスト用の質問と回答のセットデータを作成し、過去事例やガイドラインの内容に関して応答する RAG システムを開発、回答精度の検証を実施する。金融領域において、膨大な量の文書データを対象とした RAG システムとなること、文書自体に不整合や矛盾のある表現が存在する可能性が高いということという点を想定し、特に RAG における、検索用データセットの作成（Indexing）は重要であると想定される。本ペーパーでは、RAG システムの回答精度を向上させるために、Indexing においてどのような工夫が想定されるか、回答精度の向上のために有効なアプローチの検証を行う。Indexing の工夫として、1. データベース化する前に文書自体の内容を整理する、検索用の文書の分割（チャンク化）の手法における工夫、という二つの効果について検証する。データセットとして、金融庁、金融検査・監督基本方針関係[40]、監督指針・事務ガイドラインの 2 分野にてアクセスできる PDF 形式のファイル 27 件を対象ドキュメントとして検証を実施する。

データセット作成の方針としては、RAG システム精度の検証のため、クエリと回答のセット及び、回答のために必要な文書（ゴールデンドキュメント）の候補をセットとして用意を行う。これらのデータについて、RAG システムによりクエリに対する検索ドキュメント及び、回答を生成し、検索の精度に関して、LLM システムの性能評価用ライブラリ RAGAS[41]の機能を用いて、以下の方法で LLM を用いた評価を実施する。

（1）回答の適切さの評価（Answer Relevance）

目的：生成された回答が元の質問にどれだけ適切に答えているかを確認する方法

評価手順：

① LLM を使って、生成された内容が回答になるような新たな質問を作成

② 元の質問と新たに生成された質問の類似度を測定（文書をベクトル化、ベクトル間の類似度を数値的に算出する）

③ 類似度が高いほど、回答が質問に適切に答えていると判断できる

（2）文脈の精度評価（Context Precision）

目的：検索結果に質問の回答に関連する情報がどれだけ含まれているかを確認する方法

評価手順：

① 検索結果の上位のチャンクに対して、LLM を活用して各チャンクが質問に関連しているかを分析、関連性の有無を LLM に判断させる。

---

[40] 金融庁　(2024)法令・指針等　金融検査・監督方針 https://www.fsa.go.jp/common/law/index.htm

[41] RAGAS　(2024)　RAGAS Introduction　https://docs.ragas.io/en/stable/

② 関連性のあるチャンクがどれだけ含まれているかを定量的に評価

（3）応答の忠実性評価（Faithfulness）

目的：生成された回答が検索された文書の情報と整合しているかを確認する方法

評価手順：

① 生成された回答を複数の文章や文節に分割

② 分割された各文章が検索されたコンテキストから論理的に導き出せるか LLM を用いて確認、評価する

（4）文脈の再現性評価（Context Recall）

目的：生成された回答が想定される正解とどれだけ類似しているかを確認する方法

評価手順：

① 想定される正解の回答を LLM で分割

② 各部分が元の検索コンテキストと関連しているかを評価

③ 回答の内容が想定される正解とどの程度一致しているかを測定

上記 4 つの LLM を用いた定量的な評価の方法を用いて RAG システムの性能評価を行う。テストデータセットの作成の方針としては、上記の金融庁関連の PDF ドキュメントに対して LLM（OpenAI, GPT-4o）を用いて作成する。具体的にはオープンソースとして利用可能な RAG システム評価用ライブラリ（RAGAS）を用いて、検証用ドキュメントを用いてクエリと回答のセット 10 件生成を行なった。こちらのテストケースを用いて検証を実施する。特に、金融機関におけるユースケースを想定し、文書データの整理、検索、回答の生成という一連のシステムの中で特に文書データの整理のパートに関して RAG のシステムの性能への影響を評価する。金融機関においては取り扱う文書の量が多く、これらを適切に処理し、RAG システムにおける検索性能を向上させるということは重要と想定される。本ペーパーでは、既存の社内文書を分割、データベース化する際の分割の方法、及び分割を行う前の文書の不整合点、矛盾点の整理に関しての検証を実施する。作成されたテストケースのサンプルは補題 1 に追加する。サンプルの内容からも確認できる通り、指定のドキュメントを参考にして回答することが求められるような、QA 形式のデータセットが作成されている。

６.１ チャンク化（分割）

検索対象となる文書をどのような単位で分割、保存するかという点は RAG のシステムの回答精度を改善させるための工夫の余地がある。チャンク化の方法として文字数や、句読点などを基準にルールベースで分割する方法や、LLM を用いて意味合いを考慮して分割する方法もしくは人が内容を確認して意味を考慮して分割する方法が存在する。本節では、文章単位でチャンク化の方法において、チャンクの長さの変化に応じた RAG システムの回答内容、性能の変化について確認する。手法としては、指定した文字数に収まる範囲での文章単位を一つのチャンクとして分割する、という方法により RAG 検索のデータベースを構成、テストデータに対する回答を生成させる、というシナリオにて実験を行った。文字数の長さとしては 100、200、

300、400 と 4 つのパターンで検証を行い、それぞれの場合での 4 つの評価指標における値を確認した。

表 1　チャンク分割文字数と RAG 性能の評価（筆者作成）

チャンク文字数	answer relevancy	context precision	faithfullness	context recall
100	0.6088	0.3122	0.7546	0.6000
200	0.5991	0.4382	0.7630	0.6000
300	0.5842	0.3944	0.7370	0.6000
400	0.6703	0.4250	0.7216	0.6000

上記の検証結果からチャンク分割する文章の長さと RAG システムによる回答の精度の関係性について以下のことが確認できる。

（1）answer relevancy に関して、チャンク長を長くするほど、より性能が向上しているということが確認できる。この結果は、Zhao *et al.*の調査として取り上げた、正しい回答の生成のためのゴールデンドキュメントが検索結果から得られていない場合においては、入力される情報が多いほど性能が上がる、という結果と整合している可能性がある。つまり、LLM は多くの情報から回答のために必要な情報を選び出し、適切に回答をする能力があり、今回の検証においてもより長い文字数を入力した結果として回答内容が想定するものに近づいていくという傾向が認められたと考えられる。

（2）context precision に関して、検索結果のうち、回答の生成のために必要な情報がどれだけ含まれているか(どれだけノイズが少ないか)を評価する指標であるが、チャンク長の変化に対して一定の傾向は確認できていない。この結果より、今回の条件ではチャンクの長さが、クエリに対する文書の検索精度に対して明確な影響は確認されないと考えることができる。

（3）faithfulness に関して、検索結果から生成される回答の内容が導けるかどうかを評価する指標であり、チャンク文字数が長くなるほど少しずつ減少していることが確認できる。チャンクの長さが長くなるほど、回答生成時に入力される文書量も増加するということから今回の検証では LLM による評価を実施しているということもあり、LLM の評価上回答の内容と関連するコンテキストの検出における精度が下がっているということは原因と考えられる。なお、この項目に関しては、検索データの工夫だけではなく、回答生成用プロンプトの記載方法の工夫によっても大いに変化する可能性のある項目となる。本ペーパーではこちらのプロンプトエンジニアリングに関しては取り上げないが実務の上では、必要なフォーマットとして出力させるなどの実際の利用するユースケースに合わせて出力形式を制御するプロンプトエンジニアリングも必要となる。

（4）context recall の値に関しては、チャンクの長さを変更しても値に変化はなく、想定の回答を生成するために必要な文書（ゴールデンドキュメント）を検出ができているかどうかという点に関して、チャンク文字数を変化させても影響がほとんど見られなかった。

６.２ 検索対象データの最適化

RAG の回答品質を向上させるためには、検索対象になる文書の整理が重要である。金融機関において、法令対応やルールの変更などに合わせて、文書の更新をさまざまな形で実施した結果、同じ内容にも関わらず異なる表現となっている場合や内容自体に違いがあるなど、検索対象になる文章自体の不整合が多くなるというケースが存在する。このような文書自体が時間の推移によって変化し、結果として表現や内容に不整合が存在する場合に、RAG システムでは回答の精度が変化、低下してしまうということが考えられる。このことは Barnett *et al.*[42)]にても述べられている。Barnett *et al.*によれば、RAG の回答品質が下がる要因としては７つのポイントが挙げられているが、特に、以下の２つのポイントは金融機関における RAG 構築、検索システムの構築において重要なポイントと想定される。

（1）コンテンツが回答生成用プロンプトに含まれない：クエリに対する答えを持つ文書がデータベースから検索されたが、答えを生成するためのコンテキストに追加されないケース。これは、データベースから多くの文書が返され、回答を取得するために統合処理が行われた場合に発生する。

（2）コンテンツの抽出の失敗：答えはコンテキストに存在するが、回答生成時に、LLM が正しい答えを生成できないケース。通常、これはコンテキストにノイズや矛盾する情報が多すぎる場合に発生する。

（1）に記載の内容については、6.1 でも実施したチャンク方法の変更などの工夫により以下にしてクエリに対応するドキュメントを検索により抽出するか、という検索部分、Indexing における工夫が主な取り組みになる。（2）に記載のある、検索結果の中に矛盾が含まれるというケースに関して、RAG システムを開発する前段階の処理として既存の文書を LLM にて整理、矛盾点の抽出を行う、という取り組みが考えられる。矛盾の箇所の特定及び整理を実施するということを想定した場合、膨大に存在する文書を人の手で全て確認、整理するという作業は現実的ではないため、文書の不整合を LLM で検出、整理するというアプローチが有効と考えらえる。具体的なフローとしては、検索対象となる文書群の中で、類似する表現の箇所を特定する、類似箇所の表現の内容に関して、矛盾する記載が存在するかどうかを LLM もしくは自然言語処理の手法を用いて確認する、という 2 つのステップが考えられる。

検索対象となる文書群の中で、類似する表現の箇所を特定する取り組みについては、RAG の場合と同様の前処理を行う必要がある。つまり、既存の文書をチャンク化、OpenAI 社の

---

42) Scott Barnett., Stefanus Kurniawan., Srikanth Thudumu., Zach Branne, Mohamed Abdelrazek. (2024) “Seven Failure Points When Engineering a Retrieval Augmented Generation System” *CAIN 2024: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI,* pp. 194-199

Embedding-3-small などの埋め込みモデルを用いてベクトル化し、ベクトル間での類似度の評価、類似度が高いペアを抽出するという手法が考えられる。結果として表現や記載内容の類似性が高い記述箇所が発見される。

類似箇所の表現の内容に関して、矛盾する記載が存在するかどうか確認する取り組みについては、1.の方法で検出された文書ペアにおける詳細な記述内容の確認を LLM へのプロンプトとして与える、記載内容の矛盾の度合いを自然言語処理の手法を用いて評価するということが考えられる。

実際にこちらの不整合点、変化点の検出について、金融庁より公表されている「金融機関の IT ガバナンスに関する対話のための論点・プラクティスの整理」[43]の初版及び第二版[44]の 2 つ(それぞれ約 25,000 文字)を比較、不整合箇所や変化点の検出を上記フローで実施する。実施方法としては 6.1 でのケースと同様に、100 文字を分割単位としたチャンクを生成し 2 文書間での表現の類似箇所の抽出を実施した。結果は次のようになる。

表 2　2 文書の間に存在する類似表現箇所の該当箇所数（筆者作成）

	該当箇所数
類似表現箇所	127

また、この中から、2 つの文書ペアを人、もしくは LLM にて確認を行うことで不整合、不一致の表現箇所を抽出することも可能である。抽出された文書ペアのうち、表現が異なっていたケースの該当箇所について Appendix として記載をする。本アプローチを通じて、RAG システムの性能改善のために既存の文書における矛盾点を洗い出し修正の上 Indexing するということが有効な手順の一つになりうると想定され、人の手で全ての文書を確認するという工数をかけることなく対応を進めることができる可能性がある。

6 章では RAG システムの精度改善の取り組みとして、特に検索対象となる文書に対する加工に着目して検証を実施した。初めに述べたように、金融機関において RAG システムを導入するための最初のハードルとして、既存の文書量が他業界と比較し多く存在すること、また更新頻度も比較的高く、マニュアルなどの文書を検索、回答を行う RAG システムの開発、導入ハードルが比較的高いということが各種金融機関における取り組みから明らかになってきている。このような課題に対して、本ペーパーでは RAG システムの開発のために検索対象データとなる既存の文書への前処理として、不整合、矛盾表現箇所の認識と整理の実施、及び文書のチャンク化の方法の 2 点が重要なポイントとなるということを確認した。また、それぞれ実際

43) 金融庁　(2019)　金融機関の IT ガバナンスに関する対話のための論点・プラクティスの整理　初版 https://www.fsa.go.jp/news/30/20190621/01.pdf
44) 金融庁　(2023)　金融機関の IT ガバナンスに関する対話のための論点・プラクティスの整理第二版 https://www.fsa.go.jp/news/r4/sonota/20230630/02.pdf

に開発を行う上で、膨大に存在する文書の整理やチャンク化を人の手で行うというのは非現実的であると考えられる。LLM を用いることで既存の文書の整理や不整合箇所の修正を行うという取り組みの可能性を確認した。

７．結論

本ペーパーは、金融分野における LLM ユースケースを明らかにした上で、文献調査と実証分析を通じて、金融領域における LLM の評価手法の最新研究の状況と、RAG の性能、及び性能向上に向けた取り組みを検証した。金融分野において LLM を導入する際、モデルの出力は慎重な検討を要する。先行研究の調査により、金融ドメインに特化した評価指標が英語圏を中心に検討が進んでいるが、日本語の金融分野に特化した評価指標については乏しい現状が明らかになった。一方で日本でも金融に特化した LLM を作る動きは散見されており、今後金融特化型 LLM の開発と合わせて評価手法の検討も進むことが予想される。また、金融領域の LLM 活用の事例として注目されている RAG システムについて、技術的な概要から実際の導入における工夫点を整理し、実運用に向けたシステム評価方法と性能改善の方法を、金融庁のガイドライン文書などを対象に調査、評価した。具体的には、金融機関の現場で RAG システムを導入する上で必要な開発内容、精度の評価方法、及び精度向上のための取り組みについて検討した。本検証では精度評価の方法として LLM を用いたアプローチを取り上げたが、実際には人による評価も同様に重要である。しかし、膨大な量の金融機関のテキストデータを対象としたシステム開発において、全てを人手で確認・評価することは現実的ではない。そのため、LLM による初期評価を行った上で、現場のユーザーからのフィードバックをもとに改善・修正を行うアプローチが求められると考えられる。

このような開発・導入プロセスにより、ゼロから自社専用の LLM を開発するのではなく、社内専用の LLM やチャットシステムの導入が可能になると考えられる。本ペーパーは、その第一歩として、既存の文書内容を適切に整理し、Indexing するために LLM を活用することの有効性を検証した。

今後、LLM の活用がさらに進展する中で、セキュリティ対策や社内外のユーザーへの展開、そして想定されるリスクへの対応が重要な課題となるであろう。

参考文献

平野正徳 (2023)「金融分野における言語モデル性能評価のための 日本語金融ベンチマーク構築」 jxiv.564.

Araci, Dogu. (2019) “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.” arXiv:1908.10063.

Barnett, Scott., Stefanus Kurniawan, Srikanth Thudumu, Zach Branne, Mohamed Abdelrazek. (2024) “Seven Failure Points When Engineering a Retrieval Augmented Generation System.” *CAIN 2024: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI,* pp. 194-199.

Biderman, Stella., Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasil. *et al*. (2024) “Lessons from the Trenches on Reproducible Evaluation of Language Models.” arXiv:2405.14782.

Devlin, Jacob., Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2018) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of NAACL-HLT 2019,* pp. 4171-4186.

Financial Stability Board. (2024) “The Financial Stability Implications of Artificial Intelligence.” https://www.fsb.org/uploads/P14112024.pdf.

Gao, Yunfan., Yun Xiong., Xinyu Gao., Kangxiang Jia., Jinliu Pan., Yuxi Bi, *et al*. (2023) “Retrieval-Augmented Generation for Large Language Models: A Survey.” arXiv:2312.10997.

Huang, Yue., Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, *et al*. (2024) “TrustLLM: Trustworthiness in Large Language Models.” arXiv:2401.05561.

Konstantinidis, Thanos., Giorgos Iacovides, Mingxue Xu, Tony G. Constantinides, Danilo Mandic. (2024) “FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications.” arXiv:2403.12285.

Lewis, Patrick., Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, *et al*. (2020) “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems,* Article No.: 793, pp. 9459-9474.

Paperno, Denis., German Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, Raquel Fernandez. (2016) “The LAMBADA dataset: Word prediction requiring a broad discourse context.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Volume 1: Long Papers, pp. 1525-1534.

RAGAS. (2024) “RAGAS Introduction” https://docs.ragas.io/en/stable/

Rajpurkar, Pranav., Robin Jia, Percy Liang. (2018) “Know What You Don’t Know: Unanswerable Questions for SQuAD.” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* Volume 2: Short Papers, pp. 784-789.

Shah, Raj Sanjay., Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, Diyi Yang. (2022) “WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain.” *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* pp. 2322-2335.

Vaswani, Ashish., Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, Illia Polosukhin. (2017) “Attention Is All You Need.” *Advances in Neural Information Processing Systems,* 2017-December, 5999-6009.

Wang, Alex., Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. (2019) “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* pp. 353-355.

Wang, Alex., Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. (2020) “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* Article No.: 294, pp. 3266-3280.

Wang, Liang., Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei. (2024) “Multilingual E5 Text Embeddings: A Technical Report.” arXiv:2402.05672.

Xie, Qianqian., Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, *et al*. (2024) “FinBen: A Holistic Financial Benchmark for Large Language Models.” arXiv:2402.12659.

Yang, Hongyang., Xiao-Yang Liu, Christina D. Wang. (2023) “FinGPT: Open-Source Financial Large Language Models.” FinLLM at IJCAI 2023, available at SSRN: https://ssrn.com/abstract=4489826

Yu, Hao., Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. (2024) “Evaluation of Retrieval-Augmented Generation: A Survey,” arXiv:2405.07437.

Zhao, Yiyun., Hanoz Bhathena, Prateek Singh, Saket Sharma, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram. (2024) “Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain.” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 6: Industry Track*,* pp. 279-294.

Zheng, Lianmin., Wei L. Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. (2023) “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.” arXiv:2306.05685.

補論１

金融庁ガイドラインを用いて作成した RAG テストデータセットのサンプル

なお、こちらの内容は pdf のデータを元にプログラムで処理して作成したものであり、pdf 内における改行の記号(¥n)などが含まれた生のデータである。

質問：「匿名の調査シートは、どのような種類の調査で、どのような状況で使用されることを目的として配布されていますか？」

関連文書：

'新たなリスクの発生の両面からの検討を行うとともに、新たに発生する\n リスクをいかに低減・制御するかの検討を経営陣主導の下で行っている。 \n\uf06c 経営陣が企業不祥事の防止に向けて真摯に取り組んでいる姿勢を役職員に示\n すべく、経営トップと社外役員が企業不祥事の要因について議論している様子を\n 全役職員が視聴できるよう社内に発信している。 \n\uf06c 過去に発生した不祥事及びそこから得た教訓が役職員の入れ替わりと ともに風\n 化し、同種事案や根本原因を同じくする新たな不祥事が発生することのないよう、\n 経営陣が中心となって継続的な注意喚起を実施している。 \n トップダウンの指示だけではない双方向のコミュニケーションによりコンプライアン\n ス・リスク管理に資する企業文化の変革を達成すべく、経営陣と職員との意見交\n 換を定期的に実施している。 \n \n＜令和元事務年度に把握した取組み事例＞ \n「金魚鉢の金魚」という言葉を経営陣のメッセージの中心に据えることで地域金\n 融機関は、地域社会の四方八方から見られていることを経営陣が繰り返し伝え\n ることにより、役職員に対し、コンプライアンス意識の浸透を図っている。 \n コンプライアンス・リスクを軽減させるためには、経営理念を浸透させることで、役\n 職員が自身の行動が経営理念に沿ったものかどうかを常に考えるようになること\n が重要であり、そのためには役職員が真に納得できる経営理念である必要があ\n るとして、役職員皆で経営理念を見直した。 \n「□','□動指針」を具体化したものとして行動憲章を定め、「コンプライアンスの徹底」、\n「迷ったときの判断基準」を明記し、職員が自身の取るべき行動について迷ったと\n きには、法令・ルールに違反していないかのみならず、非倫理的ではないか、家\n 族に、あるいは友人に胸を張って説明できるか等の基準を示し、ルールを守って\n いれば良いという考えにはならないように努めている。 \n\uf06c 経営陣が、マネジメント実態調査を実施し、マネジメントに懸念のある拠点を把握\n し、個別に対応を行わせ たり、職員から退職の申出があった場合、表面上の理\n 由の把握にとどまらず、例えば職場におけるパワーハラスメントの有無等、その\n 背景事情を正確に把握することで、不祥事の兆候を把握しようとしている。 \n コンプライアンスに係る匿名の 調査シートを配布し、例えば、「あな たの上司は、\n コンプライアンスに則して行動をしていますか」という問 いに対して、「行動してい\n ない」という回答があった場合、臨店の際、重点的にヒアリングを行い、実態把握\n に努めている。 \n 経営理念の実現度合いを測るため、営業区域の住民を対象に、独自の幸福度\n 調査を継続的に行い、経営理念の更なる実現に向けた取組みにつなげている。 \n \n③問題事象につながった事例 \n＜平成 30 事務年度に把握した事例＞ \n 一見すると公表資料や各種研修等を通じてコンプライアンスの重要性を社内外\n に向けて発信しているように見えるものの、経営陣の根本的な姿勢は収益至上']"

回答：匿名の調査シートは、例えば「あなたの上司は、コンプライアンスに則して行動していますか」という問いに対して、「行動していない」という回答があった場合、臨店の際、重点的にヒアリングを行い、実態把握に努めている。

補論２

LLM を用いた 2 つの文書間の差分の抽出結果の一例。LLM を用いることで表現が類似しているが、内容として異なるという箇所の候補を自動で抽出することが可能である。表 3 はその一例である。

表 3　金融機関の IT ガバナンスに関する対話のための論点・プラクティスの整理、初版、第二版における典型的表現の異なる箇所の抽出結果（筆者作成）

初版	第二版
（カネの観点） （５）企業価値の創出に繋がる「IT 投資管理プロセス」IT 投資については、ROI6 等の指標を用いた事前評価・事後評価を行い、実証実験（PoC7 等）後の実用化や必要に応じてサービス自体の廃止を行うなどの PDCA を回すことが重要である	（カネの観点） （５）企業価値の創出につながる「IT 投資管理プロセス」戦略的な IT 投資額及びそれに含まれる DX 案件の投資額について、中期計画と年度予算を定め、全社的な戦略案件の起案から審議、投資意思決定までが迅速に実行できるようなプロセスを整備することが重要である

金融庁金融研究センター

〒100-8967　東京都千代田区霞ヶ関 3-2-1
中央合同庁舎 7 号館　金融庁 15 階

TEL：03-3506-6000（内線 3552）

URL：https://www.fsa.go.jp/frtc/index.html