ファイナンシャル・プランニングの自動化に向けた GPT-4 及び RAG の性能評価

西脇 一尊　大沼 俊輔　工藤 剛　門脇 一真
株式会社日本総合研究所
{nishiwaki.kazutaka, onuma.shunsuke, kudo.tsuyoshi, kadowaki.kazuma}@jri.co.jp

概要

個人の家計や人生設計について資金計画を立てるファイナンシャル・プランニングは，幅広い金融知識を必要とし，一般に専門家との相談を通じて行われる．相談を促すために，大規模言語モデル（LLM）と Retrieval-Augmented Generation（RAG）を用いた自動相談サービスの実現が期待される．しかし，LLM が RAG によって抽出された金融知識を活かして応答できるかは明らかではない．本研究ではファイナンシャル・プランニング技能検定を題材に，LLM が RAG によって抽出された金融知識を活用できるかを評価した．その結果，LLM は RAG によって得た金融知識を活用できることを確認し，自動相談サービス実現に向けた課題を明らかにした．

1　はじめに

個人の家計や人生設計について資金計画を立てるファイナンシャル・プランニングは，税制や不動産等の幅広い金融知識を必要とし，一般に専門家との相談を通じて行われる[1)]．金融機関はファイナンシャル・プランニングに関して相談できるサービスを提供しており，これを他の産業界[2)]と同様に大規模言語モデル（LLM）で自動化することができれば，顧客の利便性向上や，新たな顧客層の開拓といった効果が期待できる．LLM は，様々な分野やタスクで高い性能が報告される [1, 2, 3] 一方，誤った内容を真実のように出力する幻覚 [4] や，LLM の学習後の情報に対応できない，といった問題がある．これらの問題は，法令や年齢・金額といった数値基準の正確な理解を要するファイナンシャル・プランニングにおいては重要な課題である．

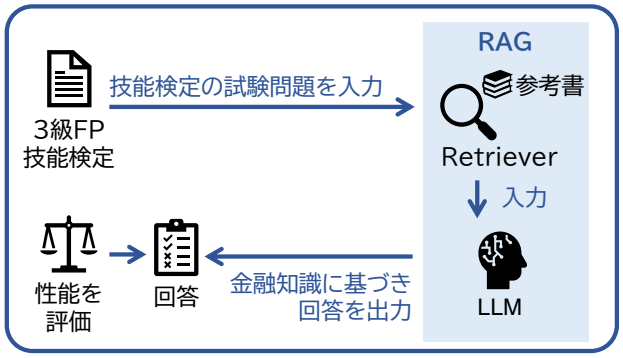

図 1　本研究における取り組みの概要．

この問題に対処する方法の 1 つに，Retrieval-Augmented Generation [5]（RAG）がある．RAG は，関連情報を含む文章を外部の情報源から抽出して LLM の入力に加えることで，その情報に基づく出力を可能にする．しかし，ファイナンシャル・プランニングの分野において，RAG を用いて LLM の性能を評価した取り組みは見当たらず，LLM が抽出された金融知識を活用できるか明らかではない．

本研究では，ファイナンシャル・プランニングの自動化に向けて，現在の LLM が RAG によって抽出された金融知識をどの程度活用できるかを明らかにする．具体的には，RAG を用いて 3 級ファイナンシャル・プランニング技能検定（FP 技能検定）に関する学習参考書のテキストを OpenAI の GPT-4 に入力し，3 級 FP 技能検定の学科試験を解かせた結果を定量・定性的に分析することで，LLM が RAG によって抽出された金融知識を活用できるかを評価する．本研究における取り組みの概要を図 1 に示す．

2　関連研究

LLM に対して金融知識に基づく出力をさせるため，金融に関連したテキストデータを用いて LLM を学習する方法 [6][7] が研究されている．一方で，RAG を用いれば，LLM の学習を行わずに金融知識

1）https://www.jafp.or.jp/aim/fptoha/fp/
2）チャット法律相談（https://chat.bengo4.com，弁護士ドットコム株式会社），Wendy's FreshAI（https://www.irwendys.com/news/news-details/2023/Wendys-Taps-Google-Cloud-to-Revolutionize-the-Drive-Thru-Experience-with-Artificial-Intelligence/，The Wendy's Company）等．

に基づいた出力が可能となると考えられる．RAG は，関連情報を含む文章を外部知識（文章集合）から抽出する Retriever と，抽出された文章を基に出力を生成する Generator の 2 つの要素からなるが，本研究では Retriever によって金融知識を抽出し，LLM に入力する．

日本語の金融分野に関するタスクにおいて，LLM と RAG を用いた取り組みが行われ，RAG は LLM に対して日本語の金融知識に基づいた出力を可能にする有効な方法であることが示されている．例えば，高野ら [8] は，投資信託の運用状況や今後の投資環境に関するコメントの生成で，市場動向に関するニュース記事を RAG で ChatGPT の入力に追加し，出力に最新の情報を反映させた．また，増田ら [9] は，GPT-4 で公認会計士試験（短答式試験監査論）を解く際に，回答に必要な知識を RAG で抽出することで，合格相当の成績を達成した．この他にも，RAG の「幻覚を抑制できる」という特徴は，日々変化する経済情勢の考慮と法令遵守が求められる金融分野において，重要な役割を果たすと考えられる．しかし，ファイナンシャル・プランニングの分野では，2 級 FP 技能検定の問題を含むベンチマークによって LLM 自体の性能を評価した試み [10] は行われている一方，RAG を用いた際の LLM の性能や課題はまだ明らかではない．そこで，本研究では LLM が RAG によって抽出された金融知識を活用できるかを評価する．

3 評価タスク

3 級 FP 技能検定とは，ファイナンシャル・プランニングに必要な技能の程度を検定する技能検定の 1 つであり，学科・実技の 2 種類の試験にて実施される．本研究では，2021 年 5 月から 2023 年 9 月までに実施された 3 級 FP 技能検定の学科試験（計 8 回分）を評価タスクとし，金融知識の他に図や表の読解が必要な実技試験は評価の対象外とした．

評価対象の試験は，いずれも 2 つの大問から構成される．各大問は，回答者に対して付随する各小問の回答方法について指示する文章が含まれている．第 1 問では，金融知識に関する文章の正誤を判定する正誤問題が，小問として 30 題出題される．第 2 問では，金融知識に関する文章中の空欄を補完する適切な文章・語句・数字，またはそれらの組み合わせを 3 つの選択肢から選ぶ三択問題が，小問として 30 題出題される．また，第 2 問において出題される小問の中には，図や表で示された情報をもとに回答をする問題もある．具体的な問題のイメージは，付録 A を参照されたい．各小問の配点は 1 題 1 点であり，合格基準は正答率 6 割以上である．実際の試験はマークシート形式で出題されている．

4 実験手順

本研究では，LLM として GPT-4 を用いた．バージョンは gpt-4-0613 を利用し，創造的な出力を避けるため temperature パラメータの値を 0 とした．評価タスクのうち，2023 年 5 月以前に実施された 3 級 FP 技能検定はリークしている可能性がある．

実験では，RAG によって金融知識が与えられた GPT-4 と，RAG を利用しない GPT-4 のそれぞれを用いて評価タスクを解かせ，LLM が RAG によって抽出された金融知識を活用できるかを評価した．

4.1 データセット構築

評価用データ　日本 FP 協会から公開[3]されている試験問題 8 回分（2021 年 5 月〜2023 年 9 月実施分）とその模範回答（いずれも PDF 形式）から，pdfminer.six[4]を用いてテキストを抽出した．この時，大問テキストに含まれる一部の表現（例：解答用紙にマークしなさい）を，プロンプトとして入力するのに適切な表現に修正した．pdfminer.six は，小問テキストに含まれる図表や，計算式中の分数を正しくパースできないため，これらを含む小問は GPT-4 への入力の対象外とし，全て誤答として評価・採点を行なった．結果，8 回分の試験問題の計 480 題から，正誤問題 240 題と三択問題 218 題をそれぞれ抽出した．

外部知識用データ　RAG で抽出する外部知識用のデータとして，3 級 FP 技能検定に関する市販の学習参考書 4 冊を用いた．参考書は OCR によって RAG で利用可能な形式（書籍情報）に変換した．

4.2 ベクターストアの構築

外部知識である文章集合に対して意味的な検索を可能にするため，書籍情報の埋め込み表現を用いたベクターストアを構築した．

まず，参考書のページ単位で書籍情報を E5 [11] の多言語版 large モデル[5]に入力し，埋め込み表現を取得した．E5 では検索クエリのテキストに *‘‘query: ’’*，

3） https://www.jafp.or.jp/exam/mohan/
4） https://pdfminersix.readthedocs.io
5） https://huggingface.co/intfloat/multilingual-e5-large

[問題]
{　大問テキスト　}
{　小問テキスト　}

1) {　選択肢 1　}
2) {　選択肢 2　}
3) {　選択肢 3　}

[参考文献]
・{　抽出された外部知識　}
...

[回答]

図 2　プロンプトの雛形．{ }はプレースホルダ．

検索対象のテキストに *‘‘passage: ’’* という接頭辞を用いることが推奨されているため，本研究でも，書籍情報の入力に *‘‘passage: ’’* という接頭辞を加えた．

ベクターストアには Chroma[6] を利用し，類似ベクトル探索は近似最近傍探索手法の HNSW [12] を，類似度はコサイン類似度をそれぞれ設定した．

4.3　外部知識の抽出

接頭辞 *‘‘query: ’’* を加えた「大問テキスト+小問テキスト」または「大問テキスト+小問テキスト+選択肢」の埋め込み表現を多言語版 E5-large で取得し，ベクターストアからの外部知識の抽出に用いた．

4.4　プロンプト設計

本研究で用いたプロンプトの雛形を図 2 に示す．なお，プロンプト中の選択肢は第 2 問の三択問題の場合にのみ含まれる．本実験では，小問 1 題につき 1 つのプロンプトを作成して GPT-4 に入力するため，1 つのプロンプトは小問テキストと，対応する大問テキストを含む．外部知識をプロンプトに含める際は，抽出結果のうち類似度について上位 5 件のテキストを，小問テキストまたは選択肢に続く位置でプロンプトに加えた．外部知識をプロンプトに含めない際には，[参考文献] の記載を省略した．

6）https://trychroma.com

4.5　評価

評価には，OpenAI Evals[7] を利用した．GPT-4 の出力結果が，模範回答と一致する場合に正答と判定した．また，三択問題については，選択肢記号だけでなく内容まで一致するか判定した．

5　結果・考察

5.1　定量評価

全体の正答率　RAG の利用有無に対する GPT-4 の正答率を表 1 に示す．表 1 より，RAG によって全体の正答率が向上し，GPT-4 の学習期間後に実施された 2023 年 9 月試験を含め各試験で合格基準を上回った．このことから，GPT-4 は RAG によって抽出された金融知識を，ファイナンシャル・プランニングに活用できると考えられる．

正誤問題における適合率・再現率　全 8 回の試験を通じた，各選択肢（①：正しい，②：誤り）の適合率と再現率を図 3 に示す．図 3 より，RAG によっていずれの選択肢においても適合率が向上したことがわかる．特に，①の適合率と②の再現率が向上したことから，RAG 無しの GPT-4 は正誤問題にて「正しい」と回答する傾向があるが，RAG によってこの傾向を緩和できたと考えられる．

三択問題における適合率・再現率　全 8 回の試験を通じた評価対象の問題について，3 つの選択肢ごとの適合率・再現率を図 4 に示す[8]．図 4 より，いずれの選択肢でも RAG によって適合率・再現率が向上したことがわかる．特に，選択肢 3 の再現率が大きく向上したことから，RAG 無しの GPT-4 は選択肢 1，2 を回答として出力する傾向があると考えられる．これは，多肢選択問題における選択肢の位置バイアス [13] による影響の可能性がある．

5.2　定性評価

定性評価では，GPT-4 の学習期間後に実施された 2023 年 9 月試験を対象に，RAG の抽出結果と GPT-4 の出力結果について分析を行なった．

RAG による知識抽出について　各プロンプトに含まれる 5 つの外部知識のうち，正答に必要な知識が 1 つ以上含まれているかを人手で評価した．正誤問題では，30 題中 29 題において，正答に必要な知

7）https://github.com/openai/evals
8）図表や数式を含む小問を除外して求めた．

表 1　評価タスクにおける正答率．**太字**は合格基準 (0.60) 以上の正答率であることを表す．

RAG	試験年月	2021/5	2021/9	2022/1	2022/5	2022/9	2023/1	2023/5	2023/9
有	第 1 問	0.80	0.83	0.77	0.73	0.77	0.90	0.77	0.87
	第 2 問	0.63	0.70	0.73	0.70	0.67	0.83	0.43	0.57
	全体	**0.72**	**0.77**	**0.75**	**0.72**	**0.72**	**0.87**	**0.60**	**0.72**
無	第 1 問	0.53	0.70	0.77	0.60	0.83	0.83	0.60	0.57
	第 2 問	0.53	0.57	0.53	0.57	0.40	0.57	0.37	0.40
	全体	0.53	**0.63**	**0.65**	0.58	**0.62**	**0.70**	0.48	0.48

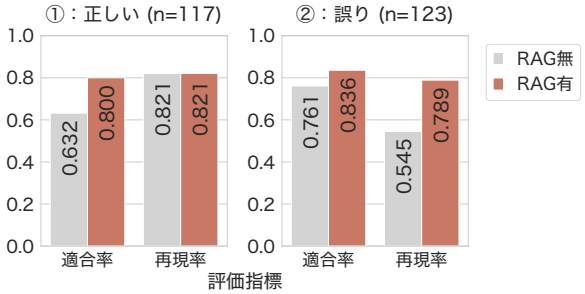

図 3　正誤問題における選択肢ごとの適合率・再現率．

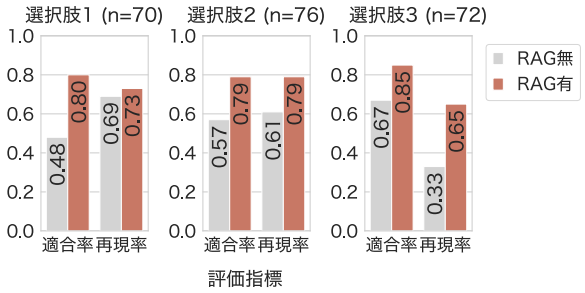

図 4　三択問題における選択肢ごとの適合率・再現率．

識を抽出できた．これは，正誤問題の文章と，書籍情報中の金融知識に関する文章が，類似していたことが原因と考えられる．

三択問題では，26 題中 20 題において，正答に必要な知識を抽出できた．抽出に失敗した問題例として，個人賠償責任保険において補償が可能な具体例を選択する問題（付録 A.1 に記載）がある．その原因として，書籍情報では「日常生活での事故による賠償に備える保険」といった抽象的な文章で個人賠償責任保険を説明している一方，選択肢で与えられた例が具体的であり，両者の類似度が低くなったことが考えられる．以上から，RAG において抽象的な知識と具体的な事象を適切に紐づける処理が，ファイナンシャル・プランニングの自動化に必要といえる．

RAG の効果について　RAG 無しの GPT-4 で誤答し，RAG によって正答した事例の一つに，2022 年 4 月から再編された東京証券取引所の株式市場区分に関する問題において，新たな株式市場区分に関する情報を抽出でき正答したと考えられる事例があった．別の事例として，所得税における特定扶養親族の年齢に関する問題において，RAG によって年齢条件（19 歳以上 23 歳未満）の正確な情報を抽出でき正答したと考えられる事例があった．

一方，同試験において RAG の有無によらず GPT-4 が誤答した一例として，選択肢に存在しない出力がされ誤答となった事例があった（付録 A.2 に記載）．他にも，RAG で正答に必要な知識が抽出できたが，GPT-4 がその内容を無視して出力したため誤答となった事例があった．いずれも，LLM が与えられた条件や外部知識に正確に基いて出力ができなかった事例であり，ファイナンシャル・プランニングの自動化に向けて解決が必要といえる．

6　おわりに

本研究では，ファイナンシャル・プランニングの自動化に向けた第一歩として，LLM が RAG によって抽出された金融知識を活用できることを確認するとともに，ファイナンシャル・プランニングの自動化に向けたいくつかの課題を明らかにした．

定量評価によって，GPT-4 に対して RAG を用いて金融知識を入力することで，3 級 FP 技能検定の学科試験における正答率が向上し，LLM が RAG によって抽出された金融知識を活用できることを確認した．定性評価によって，ファイナンシャル・プランニングの自動化には，RAG において抽象的な知識と具体的な事象を紐づける処理と，入力された条件や外部知識を LLM の出力に正確に反映させる処理の実現が必要と確認した．

今後は，ファイナンシャル・プランニングの自動化に向けて，より難易度の高い FP 技能検定へ取り組むとともに，図表・数式を処理する機能，相談内容に応じた判断を可能とするエージェント機能の実現にも取り組む予定である．

参考文献

[1] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. **Nature**, Vol. 620, No. 7972, pp. 172–180, 2023.

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. **arXiv preprint arXiv:2204.01691**, 2022.

[4] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On Faithfulness and Factuality in Abstractive Summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, 2020. Association for Computational Linguistics.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[6] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance. **arXiv preprint arXiv:2303.17564**, 2023.

[7] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models. **FinLLM Symposium at IJCAI 2023**, 2023.

[8] 高野海斗, 中川慧, 藤本悠吾. ChatGPT を活用した運用報告書の市況コメントの自動生成. 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-031, pp. 61–67, 2023.

[9] 増田樹, 中川慧, 星野崇宏. ChatGPT は公認会計士試験を突破できるか？: 短答式試験監査論への挑戦. 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-031, pp. 81–88, 2023.

[10] 平野正徳. 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 preprint jxiv.564**, 2023.

[11] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training. **arXiv preprint arXiv:2212.03533**, 2022.

[12] Yu A. Malkov and D. A. Yashunin. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 42, No. 4, pp. 824–836, 2020.

[13] Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. **arXiv preprint arXiv:2308.11483**, 2023.

A　付録

A.1　RAG による知識抽出に失敗した問題の例

選択肢で与えられた具体的な事例と，書籍情報中の個人賠償責任保険についての説明文の紐付けができず，RAG に失敗したと考えられる例を図 5 に示す．なお，この問題の正答は「3) 買い物中に誤って商品を落として破損させてしまい」である．

[問題]
次の文章の（　　）内にあてはまる最も適切な文章，語句，数字またはそれらの組合せを 1)〜3) のなかから選び，その番号を解答しなさい．
個人賠償責任保険（特約）では，被保険者が（　　），法律上の損害賠償責任を負うことによって被る損害は，補償の対象となる．

1) 業務中に自転車で歩行者に衝突してケガをさせてしまい
2) 自動車を駐車する際に誤って隣の自動車に傷を付けてしまい
3) 買い物中に誤って商品を落として破損させてしまい

図 5　RAG による知識抽出に失敗した問題の例．問題文の出典：金融財政事情研究会 / 日本ファイナンシャル・プランナーズ協会 ファイナンシャル・プランニング技能検定 3 級学科試験（2023 年 9 月）を一部改変．

A.2　選択肢に存在しない出力がされた例

選択肢に存在しない回答が出力され，誤答となった例を図 6 に示す．なお，この問題の正答は「2) ① 贈与税 ② 80%」である．

[問題]
次の文章の（　　）内にあてはまる最も適切な文章，語句，数字またはそれらの組合せを 1)〜3) のなかから選び，その番号を解答しなさい．
相続税路線価は，相続税や（　①　）を算定する際の土地等の評価額の基準となる価格であり，地価公示法による公示価格の（　②　）を価格水準の目安として設定される．

1) ① 贈与税 ② 70%
2) ① 贈与税 ② 80%
3) ① 固定資産税 ② 80%

[参考文献]
（省略）

[回答]
1) ① 贈与税 ② 80%

図 6　選択肢に存在しない回答が出力された例．[回答] の後に続く赤字は GPT-4 の出力結果を表す．問題文の出典：金融財政事情研究会 / 日本ファイナンシャル・プランナーズ協会 ファイナンシャル・プランニング技能検定 3 級学科試験（2023 年 9 月）を一部改変．