保険の2030年アジェンダ
シリーズ23

保険会社における生成AI活用実践（育成編）

生成AIは、単なる業務効率化に留まらず、保険会社社員と同等レベルでの業務実行可能性を秘めている。前号の採用編では、生成AIの素のモデル（Pre-trained）の評価を行った。今号では選定された生成AIを育成することで、学科試験（FP3級）とオリジナルの実技試験の結果がどの程度向上するか評価を行った。

生成AI育成手法としてのRAGとFT

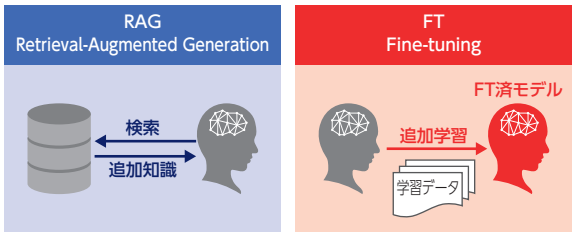
生成AIの育成手法としてはRetrieval-Augmented Generation（RAG）とFine-tuning（FT）がある。

RAGは生成AIのモデルには手を加えないものの、外部データの検索を行い追加で知識を与えることで、より専門的な回答を得る手法である。FTは追加で学習を行い、推論モデル（人でいう思考回路）自体を調整することでコンテキストに応じた回答が可能になる（図表1）。

FTは2023年8月に公開されたばかり[1]であり、現在の主流はRAGでFTの事例はまだ少ない。これはRAGの方が検索されたデータをもとに推論するため回答根拠を確認できる一方で、FTは回答根拠がブラックボックスで評価しにくいためである。今回、RAGとFTによる育成の結果を比較することで、生成AIを新たに業務活用する際の参考になると考えた。

前号の採用編で選定した3つのモデルのうち、2023年12月時点でAPI（アプリケーション・プログラミング・インターフェース）が公開されているGPT-3.5のRAGとFT、GPT-4のRAGを対象に育成を行った[2]。GPT-4のFTは開発中のため非公開で育成できなかった。

図表1　生成AIの育成手法



（出所）野村総合研究所

予想と反する育成結果

両手法の特徴から、RAGにより学科試験（FP3級）の正答率が向上し、FTにより実技試験の得点率が向上すると予想した。しかし、育成結果は予想と異なった。

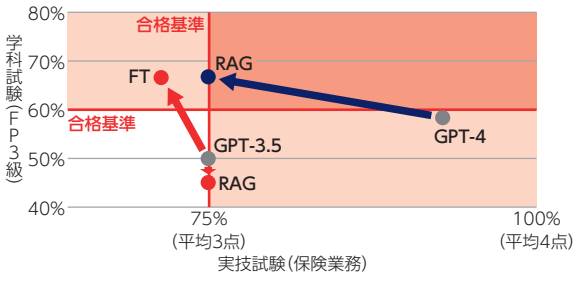
RAGによりGPT-3.5の学科試験の正答率は下がり、GPT-4でもわずかな上昇となった。FTにより、GPT-3.5の実技試験の得点率は下がったが、学科試験の正答率が大幅に上がった（図表2）。

RAGで予想よりも学科試験の正答率が上がらなかった原因の一つに、エンベディング[3]モデルがあると考える。今回はOpenAIの「text-embedding-ada-002」を用いたが、新しい高効率モデルも発表されており[4]、モデルの選択により結果が変わる可能性がある。

FTで実技試験の総合得点は下がったが、個々に見ると7問中4問では得点が上がり、3問では下がっている。得点を下げた理由としては、学習データ不足とハルシネーション[5]の問題と考える。

今回、学習データにはFP3級の過去問を用いたが、実技試験の中にはライフプランニングを行う設問があり、FP3級の知識だけでは不十分だったと推測する。CFP

図表2　育成後の試験結果



（出所）野村総合研究所

NOTE

1) 2023年8月以前にもGPT-3ベースのモデル向けのFTは公開されていたが、GPT-3.5向けのFTが公開された。https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates
2) RAGとFTの育成用データには共通でFP3級の過去問を2016年5月から2022年1月まで17回分1020問用意した。いずれもOpenAIが提供するAPI（RAGにはhttps://api.openai.com/v1/embeddings、FTにはhttps://api.openai.com/v1/fine_tuning/jobs）を使用して育成を行った。
3) テキストをベクトル化する手法。これにより類似度の高いデータの検索が可能になる。
4) 2024年1月に2つの新しいエンベディングモデル「text-embedding-3-small」「text-embedding-3-large」が公開された。https://openai.com/blog/new-embedding-models-and-api-updates
5) AIが事実に基づかない情報を生成する現象のこと。まるでAIが幻覚（＝ハルシネーション）を見ているかのように、もっともらしい嘘（事実とは異なる内容）を出力するため、このように呼ばれている。https://www.nri.com/jp/knowledge/glossary/lst/ha/hallucination
6) 形式変換では、穴埋め問題において穴埋め箇所を括弧から記号に変換するとともに、複数穴埋め箇所がある場合は問題をそれぞれに分けた。

（認定ファイナンシャル・プランナー）など、より専門的な知識を追加で学習させることにより改善を見込む。

ハルシネーションについては、例えば資産形成への意識が高い20代女性向けの保険提案問題に対して「インデックスリンクプラン」の終身保険など、存在しない保険商品の提案をしてしまった。これは、FP3級の過去問にあった「インデックス型投資信託」などの単語が推論モデルに影響を与えてしまい、設問の「資産形成」という単語から類推されたものと考える。

得点が上がったのは、「病気やけが、認知症、介護が必要になった場合に心配だ」など比較的簡易な設問で、FP3級の過去問により保険や特約について学習した効果があったと考える。

FTのポイントと可能性

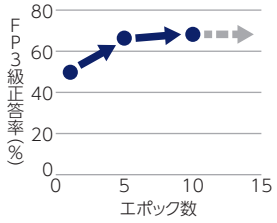
FTによる育成でポイントとなるのは学習データとエポック数である。

今回はFP3級の過去問を学習データとしたが、実際には業務マニュアルや保険営業のトークスクリプトなど、より実践的なデータであることが望ましい。ただし、学習データは生成AIが理解できるようなQA形式である必要がある。今回は過去問からの形式変換[6]で70時間費やしたが、業務マニュアルなどの非構造データでは、より変換にかかる工数が多くなると考えられる。

エポック数とは生成AIのモデルが学習データを学習する回数のことで、回数が少ないと学習が終わらず回数が多いと過学習になるため、適切な数を見つける必要がある。今回、エポック数を変えて試したところ図表3のように、エポック数5回と10回ではFP3級の正答率はあまり変わらず、コストが倍になる結果となったため、エポック数は5回とした。

図表3　エポック数の比較

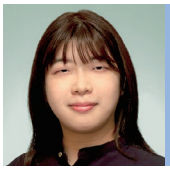
エポック数	FP3級正答率（%）	OpenAI API利用料（$）
1	50.0	3.3
5	66.6	6.6
10	68.3	13



（出所）野村総合研究所

今回、FTで学科試験の結果が大幅に上昇した。これは、定型的な質問回答はFTよりRAGが適しているという予想を裏切るものであった。さらに、FTは実技試験でも7問中4問の得点が上がっている。課題はハルシネーションだが、今回の発生箇所は保険商品の名称に限定されており、商品パンフレットを参照させるなどの対策で回避できる可能性がある。

育成後の生成AIも単独での活用はまだ現実的でなく、例えば保険提案におけるベース案作成など営業補助ツールとしての活用が有効だと考える。しかし、それが生成AIの限界というわけではない。今後も継続的に生成AIを育成することで、本当に保険営業職員の相棒になるようなAIを生み出したい。今回、実際に育成に着手し、生成AIが人の相棒になる未来は5年後10年後ではなく、すぐそこにあると感じた。



Writer's Profile
虎瀬 なつみ Natsumi Torase
保険デジタル企画部　未来保険研究室
システムコンサルタント
専門はAIとスクラム開発
focus@nri.co.jp