Generative AI use in insurance industry (model training)

Natsumi Torase

18 March 2024

Iakyara vol.383

Nomura Research Institute, Ltd.

Natsumi Torase
System consultant
Digital Insurance Planning Department

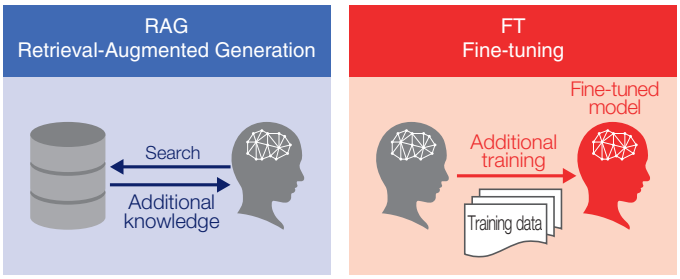
Executive Summary

Generative AI has the potential to not only improve insurers' operating efficiency but also perform insurance industry jobs as well as humans. In lakyara vol. 382, we evaluated pre-trained generative AI models to select one for an AI development program for the insurance industry. In this installment, we evaluate the extent to which training improved the selected models' performance on an academic test (level-3 financial planning certification exam) and a practical skills test we devised in house.

Training generative AI models with RAG and FT

Generative AI training techniques include retrieval-augmented generation (RAG) and fine-tuning (FT). RAG does not involve modification of the AI model itself. Instead, RAG uses external data retrieval to enable the model to answer queries more expertly by making more knowledge available to it. FT enables AI models to generate contextually accurate answers by tuning the AI inference model (analogous to human thought processes) through additional training (Figure 1).

Figure 1. Generative AI training techniques



Source: NRI

NOTE

1) FT became available for GPT-3.5 in August 2023. It was previously available for GPT-3 base models. See https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates.

FT is newer than RAG. It first became available in August 2023[1] and consequently has not been used as much in practice as RAG, currently the most common training technique. With RAG, the way in which the model infers answers from retrieved data is transparent. With FT, by contrast, such inference is more of a black box and therefore harder to evaluate. In our project, we trained models with RAG and FT and compared the results to shed light on new business use cases for generative AI.

In *lakyara vol. 382*, we selected AI models as candidates for an AI development

©2024 Nomura Research Institute, Ltd. All Rights Reserved.

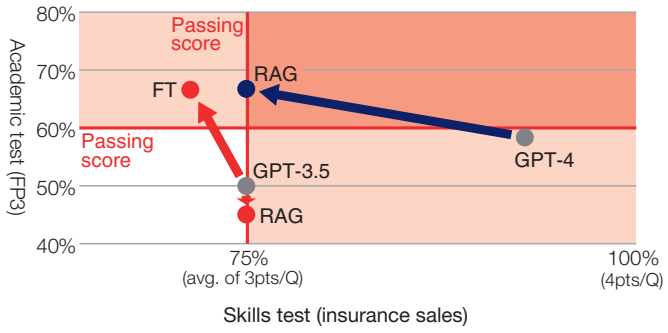
program for the insurance industry. Of the chosen models, we trained GPT-3.5 using both RAG and FT and GPT-4 using RAG alone because its API did not offer FT at that time (December 2023)[2] .

2) We used level-3 financial planner certification exam questions as both RAG and FT training data. Specifically, we used 1,020 questions from 17 exams administered between May 2016 and January 2022. We trained the models with RAG (https://api.openai.com/v1/embeddings) and FT (https://api.openai.com/v1/fine_tuning/jobs) via OpenAI APIs.

Unexpected training results

Given how RAG and FT work, we expected RAG to increase the models' scores on the academic test (Japanese level-3 financial planner (FP) certification exam) and FT to increase GPT-3.5's skills test score. In actuality, however, the training results defied our expectations. RAG's effect on academic test scores was negative for GPT-3.5 and barely positive for GPT-4 while FT decreased GPT-3.5's skills test score but substantially improved its academic test score (Figure 2).

Figure 2. Post-training test results



Source: NRI

RAG's failure to increase academic test scores as much as we had expected may be partly due to the embedding[3] model we used: OpenAI's text-embedding-ada-002. OpenAI has since released upgraded embedding models[4]. RAG may yield better results if a different embedding model were used.

3) Embeddings are vectorized representations of text strings. They enable retrieval of data highly relevant to the query.

4) OpenAI released two new embedding models in January 2024: text-embedding-3-small and text-embedding-3-large. See https://openai.com/blog/new-embedding-models-and-api-updates.

While FT resulted in a lower overall score on the skills test, the fine-tuned GPT-3.5 model scored higher on four of the test's seven problems and lower on the other three. We attribute the lowered scores to inadequate training data and hallucination[5].

5) Hallucination is AI-generated false information.

For training data, we used questions and answers from past level-3 FP certification exams. Our skills test, however, included problems that involved life planning, a discipline not adequately covered by the level-3 FP curriculum. The model's skills test performance would likely improve with additional training on more specialized topics covered by, e.g., the CFP exam curriculum.

©2024 Nomura Research Institute, Ltd. All Rights Reserved.

The hallucination took the form of recommendations of nonexistent insurance products. For example, given the task of providing insurance recommendations to a woman in her 20s who was interested in building wealth, the model recommended “index-linked whole-life insurance,” presumably because it had associated “building wealth” with index-linked investment trusts, which were mentioned in exam questions used as training data.

The skills test problems on which FT had improved the model’s scores were relatively easy ones, including a question on contingency planning for illness, injury, cognitive impairment and long-term care. We attribute the improved scores on these problems to the additional past FP exams we used to train the model on policy riders and insurance in general.

FT pointers and possibilities

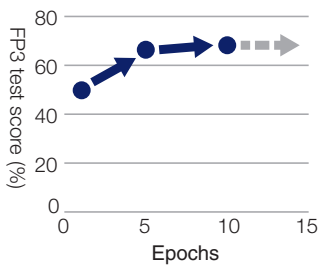
The key determinants of FT’s effectiveness are the training data’ relevance and the number of training epochs. Although we used past FP exams as our training data, we actually should have used more practical data like operations manuals and insurance sales presentation scripts. However, training data needs to be in Q&A format to be comprehensible to generative AI models. We spent 70 person-hours reformatting exam questions[6]. This process would likely have been even more labor-intensive if we had used unstructured data like operations manuals.

6) Reformatting involves replacing the blanks in fill-in-the-blank questions with symbols and converting questions with multiple blanks into separate questions with one blank apiece.

The number of epochs is the number of times that training data is fed through a generative AI model to train it. Too few epochs leave the model insufficiently trained while too many epochs result in overtraining. It is therefore necessary to ascertain the optimal number of epochs. Figure 3 shows how model performance and training costs vary as a function of the number of epochs. Model performance

Figure 3: Model performance and training cost by number of epochs

Epochs	FP3 test score (%)	OpenAI API usage charges ($)
1	50.0	3.3
5	66.6	6.6
10	68.3	13



Source: NRI

©2024 Nomura Research Institute, Ltd. All Rights Reserved.

as measured by the percentage of level-3 FP exam questions answered correctly was nearly unchanged between five and 10 epochs while the cost roughly doubled. We accordingly decided to do five epochs of FT.

FT substantially improved model performance on the academic test, contrary to our expectation that RAG would be better-suited to training a model to answer formulaic questions than FT would. Additionally, FT improved the model's scores on four of the skills test's seven problems as well. While hallucination was an issue, the hallucinated output was limited to insurance product names. We may be able to avoid hallucination by augmenting the training data with, e.g., product pamphlets.

Even after being trained, generative AI models are not yet realistically ready to be deployed on a solo basis, albeit not because of any shortcoming of generative AI itself. We believe generative AI today can effectively assist insurance sales reps by, for example, generating first drafts of insurance proposals. We plan to continue working on developing generative AI models capable of truly partnering with insurance sales reps. Now that we have started to actually do so, we believe such models will be a reality not in five or 10 years but much sooner.

©2024 Nomura Research Institute, Ltd. All Rights Reserved.

about NRI

Founded in 1965, Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services with annual sales above $5.1 billion. NRI offers clients holistic support of all aspects of operations from back- to front-office, with NRI's research expertise and innovative solutions as well as understanding of operational challenges faced by financial services firms. The clients include broker-dealers, asset managers, banks and insurance providers. NRI has its offices globally including New York, London, Tokyo, Hong Kong and Singapore, and over 16,500 employees.

For more information, visit https://www.nri.com/en

The entire content of this report is subject to copyright with all rights reserved.
The report is provided solely for informational purposes for our UK and USA readers and is not to be construed as providing advice, recommendations, endorsements, representations or warranties of any kind whatsoever.
Whilst every effort has been taken to ensure the accuracy of the information, NRI shall have no liability for any loss or damage arising directly or indirectly from the use of the information contained in this report.
Reproduction in whole or in part use for any public purpose is permitted only with the prior written approval of Nomura Research Institute, Ltd.

Inquiries to : Financial Market & Digital Business Research Department
Nomura Research Institute, Ltd.
Otemachi Financial City Grand Cube,
1-9-2 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan
E-mail : kyara@nri.co.jp

https://www.nri.com/en/knowledge/publication/fis/lakyara/

©2024 Nomura Research Institute, Ltd. All Rights Reserved.