NRI

Generative AI use in insurance industry (model selection)

Hiroko Washiyama

11 March 2024

lakyara vol.382

Nomura Research Institute, Ltd.

Executive Summary

Generative AI has the potential to not only improve insurers' operating efficiency but also perform insurance industry jobs as well as humans. We designed an AI development program for insurers and evaluated five generative AI models' applicability to the insurance industry.

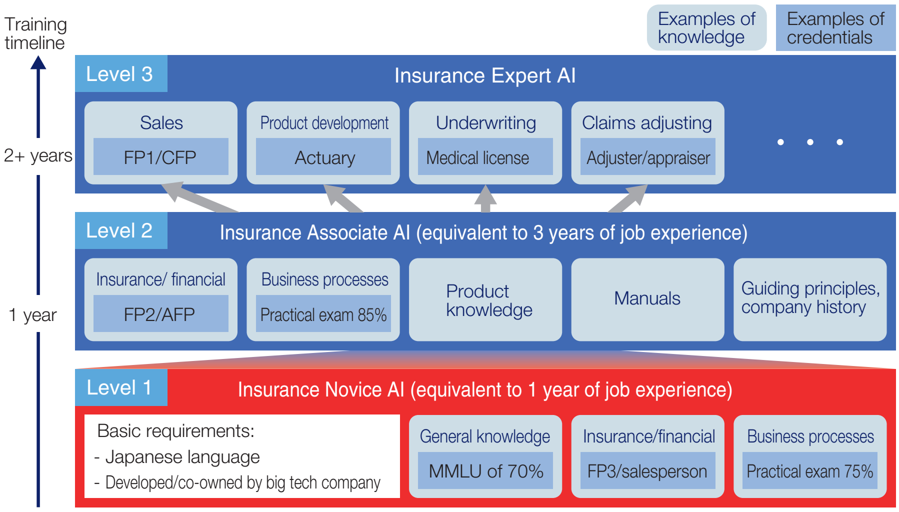
Hiroko Washiyama
Expert consultant
Digital Insurance Planning Department

Can AI perform an entry-level job in insurance industry?

In the near future, insurers may be utilizing multiple specialized AIs just as they now employ humans with different types of expertise. Insurers' AI use to date has been mainly limited to technologies like AI-OCR that improve clerical processes' output quality and efficiency. With the advent of generative AI capable of natural language interactions, AI now has the potential to take over even creative work.

In anticipation of such a future, NRI has devised an AI development program modeled after insurers' career development programs for their staff (Figure 1). The program comprises three levels. We have named the levels, in ascending order of specialization, the Insurance Novice AI, Insurance Associate AI and Insurance Expert AI. The Insurance Novice AI is endowed with knowledge and skills equivalent to a typical employee with one year of insurance industry experience. The Insurance Associate AI will be equivalent in knowledge and skills to a typical employee with three years of experience. In addition to being more

Figure 1. AI development program for insurers



Source: NRI

specialized than the Novice AI, the Associate AI will also possess company-specific knowledge of products and even the company's guiding principles. The Insurance Expert AI will have even more insurance business expertise. It will require continuous feedback and follow-up training. We envision it being trainable by other AIs in addition to AI engineers.

The next step after formulating this AI development program was to deploy (select) a generative AI model from among the various ones available, taking into account their respective strengths and performance. The model evaluation and selection process we used for the Insurance Novice AI is described below. We elected to use a pre-trained model with no follow-up training. Model training is discussed in *lakyara vol. 383.*

Evaluation criteria for selecting Insurance Novice AI model

First, we identified generative AI model candidates based largely on our project's basic requirements. Specifically, we wanted an AI model with Japanese language capabilities and an MMLU score[1], a benchmark of AI models' general comprehension level, of at least 70%. Given the prospects of continued innovation in the AI space, we also wanted a model developed or at least co-owned by a big tech company. Using these criteria, we narrowed down the available AI models to five candidates: GPT-4, GPT-3.5, PaLM2, Claude2 and Gemini Pro[2]. In an HR context, this step would be analogous to screening job applicants by vetting their resumes and having them take an aptitude test. Unlike in a hiring process, we may need to also assess AI models' political biases[3] but we chose to address this issue at a later date.

Next, we subjected the candidate models to academic and skills testing. For the academic test, we used 60 questions taken from a Japanese level-3 financial planner (FP) certification exam administered in May 2023. We chose this exam because many insurers offer not only insurance but other financial products as well. We set the passing score at 60%, the same score required for human applicants to pass the level-3 FP certification exam. Additionally, a 60% score on the exam coincides with the level of financial and insurance knowledge that insurers' employees should have after one year on the job.

For the skills test, we formulated questions about insurance sales scenarios because insurers generally require their employees to be licensed as insurance

NOTE

[1] MMLU (massive multitask language understanding) score is a benchmark designed to assess natural language models' knowledge and problem-solving capabilities. It covers 57 subjects including elementary mathematics, history, computer science and law.

[2] We used the gpt-3.5-turbo-0613 version of GPT-3.5, text-bison@001 version of PaLM2 and gemini-pro version of Gemini Pro. For GPT-4 and Claude2, we used the latest publicly available browser versions as of November 2023. We used a temperature setting of zero for GPT-3.5, PaLM2 and Gemini Pro and the default temperature settings for GPT-4 and Claude2.

[3] See Shangbin Feng et al., *From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models* (July 2023).

Figure 2. Sample test question: AI-generated insurance recommendations for woman in her 20s

		GPT-4	GPT-3.5	Claude2	PaLM2	Gemini Pro
Score (scale of 1 to 4)		4	4	3	3	3
Justification for score		Recommended products for both risk mitigation and wealth building	Recommended products for both risk mitigation and wealth building	Recommended products for risk mitigation only	Recommended products for risk mitigation only	No difference in recommendations for single and partnered
Partnered	Risk mitigation	Life insurance Medical insurance Educational insurance	Life insurance Medical insurance Educational insurance	Life insurance Medical insurance Credit life insurance	Life insurance Educational insurance	Life insurance Medical insurance Cancer insurance
	Wealth building	Universal life insurance	Individual annuity			Universal life insurance
Single	Risk mitigation	Life insurance Medical insurance Disability insurance	Medical insurance	Long-term care insurance	Medical insurance Cancer insurance	Life insurance Medical insurance Cancer insurance
	Wealth building	Universal life insurance	Individual annuity			Universal life insurance

Source: NRI

salespersons. The AI models' answers were scored on a 4-point scale by NRI colleagues with insurance backgrounds.

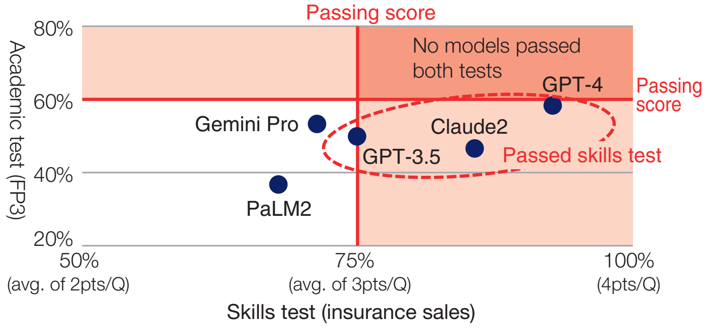
As an example, Figure 2 shows how the AI models answered one of the skills test questions. The question asked the models to recommend insurance as an outside salesperson to a woman in her 20s who lives in Tokyo, has an annual income of ¥8mn and is interested in building wealth. The models were instructed to provide two sets of recommendations, one that assumed the woman is single and other that assumed she is married or otherwise partnered. To earn a score of 4 points, the models had to recommend suitable insurance products that address both risk mitigation and wealth building under each of the two relationship-status assumptions. Recommendations that failed either to differentiate between the two relationship-status scenarios or to address wealth-building received 3 points; recommendations that were questionable in terms of product suitability received 2 points; and unsuitable recommendations received 1 point. To be eligible to be selected as the Insurance Novice AI model, candidate models had to score at least 75% (average of 3 points per question) on the skills test.

Pre-trained model evaluation results

None of the generative AI models scored high enough on the academic test (level-3 FP certification exam) to qualify as the Insurance Novice AI model, but GPT-4, Claude2 and GPT-3.5 all passed the skills test with scores of at least 75% (3 points per question on average), as shown in Figure 3. The best performer was GPT-4 with scores of 58.3% on the academic test and 93% on the skills test. We chose the three models that passed the skills test to advance to the next phase of our project because we were confident that, with additional training, they would pass the academic test. The process by which we raised their academic test scores through additional training is described in *lakyara vol. 383.*

While we used level-3 FP certification exam questions and insurance sales scenario questions to evaluate AI models, insurers would likely each make different choices in terms of test content and scoring. With new generative AI models now being developed in rapid succession, insurers need to set their own evaluation criteria and constantly keep abreast of new models' potential to replace existing ones.

Figure 3. Pre-trained model test results



Source: NRI

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 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.

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/