AI solutions for advanced enterprise needs

- Interview with Saurabh Baji
by Keiko Mukai -

4 June 2025

Special Edition

lakyara vol.400

Nomura Research Institute, Ltd.

Executive Summary

In the financial industry, the adoption of large language models (LLMs) and other AI technologies in business operations is accelerating. However, challenges remain—particularly around ensuring the safety and reliability of these tools. What kinds of LLM solutions can truly meet the high standards required by enterprise users? To explore this, we spoke with Saurabh Baji, a key figure in technological development at Cohere, a leading AI company co-headquartered in Toronto and San Francisco providing state-of-the-art LLMs to enterprise clients.

Saurabh Baji

CTO
Cohere

Saurabh Baji is CTO of Cohere. His responsibilities include machine learning, engineering and product management. Prior to Cohere, Saurabh held leadership positions over 20 years at AWS, Unity and Quantcast, building and operating multi-billion dollar products globally with AI across domains like Big Data, Robotics, Scientific Computing, Logistics, Supply Chain Optimization and more.

Keiko Mukai

General Manager
Financial AI Platform Promotion Department,
Nomura Research Institute, Ltd.

Keiko Mukai has been leading the planning of AI-driven platforms for financial institutions in her current role at NRI since 2025. She was appointed Head of the Securities Project Planning Office in 2016. Before that, she was responsible for planning and developing front-office systems for major securities firms and managing online trading system businesses. She joined NRI in 1999, following the start of her career at a major IT company in 1997.

The strengths of Cohere's generative AI

Mukai: Cohere provides large language models (LLMs) for enterprise clients. Could you start by giving us an overview of your business and your core offerings?

Baji: Cohere is a company that provides both AI models and comprehensive solutions to enterprises. Underlying what we do is our philosophy that generative AI isn't just a convenient tool, it's something that can solve the problems that companies are facing.

In recent years, the field of generative AI has seen almost magical technological innovations, which have made it possible to easily accomplish tasks that AI had previously been ill-equipped to handle. However, even as AI models have continued to evolve, most companies haven't managed to leverage them very well for their actual businesses. Having seen this situation firsthand, we chose to devote ourselves not only to model development, but also to providing solutions that enable AI to be deployed in actual business operations.

At Cohere, we provide several different generative AI models customized to our clients' needs such as Command A, R, R+, R7B. These models come equipped with various multilingual features, like generating or summarizing text, retrieval augmented generation (RAG), which is used to implement FAQs using in-house information, and more.

Regarding RAG, there are three critical components that determine its performance. The first is the accuracy of search and retrieval. This pertains to how precisely relevant information can be located and extracted from data sources in response to a user's query. The second component is the relevance evaluation of the retrieved information. It is crucial to be able to to sift through the multiple pieces of information retrieved during the search and select the most pertinent and significant items that directly align with the intent behind the user's question. The third element is the LLM's capability in executing the RAG process. It is important that the generative AI model can effectively manage tasks such as decomposing and reformulating search queries (sub-query generation) to obtain more precise information, as well as integrating and handling information gathered from multiple, distinct data sources. Beyond that, we also provide search and retrieval models.

These include technologies for enhancing search accuracy (embedding) and technologies for rearranging search results based on relevance (reranking). These models help to accurately and securely extract related information from the vast amounts of data companies possess. Given that companies hold tremendous volumes of data, the quality of these models significantly contributes to RAG's overall capabilities.

Mukai: I understand that the reranking model you mentioned is one of your company's strong suits.

Baji: Yes. Our reranking model is industry-leading, and we were actually the first to bring this capability to market.

Reranking helps enterprises significantly reduce the cost of AI deployment. Companies possess a huge volume of documents, and when they try to process all of them as input data in AI models, it gets extremely expensive. However, if they use embeddings to do a semantic search (a search based on the meaning or context of a query), and then perform reranking on the results, they can instead transfer only the most highly relevant data over.

Mukai: So reranking helps control costs by narrowing down the data before feeding it into the LLM?

Baji: That's right. To give you an example—if you know the answer you're looking for is on a single page of a book, there's no need to input the entire book. Using embeddings, you can narrow down the content to about ten potentially relevant pages. Then, with reranking, the model identifies the most relevant page, and only that page is passed to the generative model.

As one more example, what our generative models excel at is generating answers with verifiable citations attached to the data that's provided by the AI's response. Cohere has been providing this citation feature for about a year and a half now.

Mukai: So you're able to trace each part of a generated answer back to a specific page in a specific document?

Baji: Exactly. Hallucination—where the AI generates non-factual information—is a major concern for enterprises. With citations, users can verify where the information comes from, significantly reducing the risk of hallucinations. And when the information simply doesn't exist, the model will say "no answer"—which is certainly better than giving the wrong one.

How can companies' high-level needs be met?

Mukai: Japanese financial institutions are known for keeping large volumes of documents, including many that are outdated due to system or regulatory changes. A human could judge which information is obsolete, but AI might mistakenly give high relevance to such outdated content. How does your reranking model handle this?

Baji: Reranking determines relevance by evaluating a variety of factors—what we call "relevance scores." Among them, temporal signals, which indicate how recent the information is, are among the most critical. Just as you pointed out, even if an old document matches the query perfectly, it may no longer be relevant to current business needs.

Additionally, temporal constraints can be included in the query itself. Our most recent reranking model, released last December, is much more capable of inferring these constraints. It can understand requests like "show only results from the past month," and respond accordingly.

Mukai: Can users configure which factors are prioritized in the reranking process, or is that automatically managed?

Baji: Our reranking model has been trained with a wide range of use cases in mind. By default, it automatically balances various relevance scores—such as recency, content relevance, and so on—to produce the best result. Basically, the AI will automatically make the optimal determination, but depending on the customer's needs, it can also be adjusted to emphasize particular factors.

Mukai: When it comes to LLMs for enterprise use, security and privacy are key concerns. What steps is Cohere taking to address these?

Baji: We've made it possible for our models and solutions to be deployed no matter where our customers' data is being kept. We provide SaaS solutions, just as our competitors do, but ours can also be used in managed services from major cloud providers like OCI, AWS, and Azure. What's more, the exact same technology can also be deployed in virtual private clouds, on-premise environments, and air-gapped environments.

This means that even for high-confidentiality use cases, customers can deploy our models securely, without worrying about data leakage or exposure.

Handling generative AI's various risks

Mukai: Europe and the U.S. seem to approach AI regulation very differently—Europe leans toward strict regulations out of caution, while the U.S. promotes innovation and open development. Which approach is more aligned with Cohere's perspective?

Baji: We recognize that the evident and practical risks of AI like security and privacy are problems that concern our customers. That's why we spare no effort in helping our customers to minimize those risks. Conversely, when it comes to long-term threats posed by AI to humanity's existence, for example, we think it's unlikely that such risks will actually materialize. Various industries and governmental organizations are keeping an eye on those kinds of risks, and we believe that if any alarming real-world problems do arise, we'll be able to handle them.

We believe that if technology is constrained too much, its benefits can no longer be fully enjoyed, and so I think we're closer to the US approach. That said, we also understand clearly that the perceptions of the risks vary from one region to another, which is why we work together with our customers in appropriately responding to the regulations in each country and region.

Mukai: Japanese financial institutions are particularly concerned about AI infringing on human rights. How do you address that?

Baji: That's a very important issue. In most cases, companies are looking for the safest possible answers, and so our model was designed in its default configuration to provide extremely conservative responses. However, with something like red teaming (a type of vulnerability assessment), where risky answers are deliberately sought, the safety mode can be disabled, thereby allowing the AI to respond more flexibly.

Mukai: In recent years, it's become increasingly important for Japanese financial institutions to take measures involving economic security. Is Cohere also considering any measures for legal or regulatory compliance in terms of economic security?

Baji: We have put preventive measures in place to address such problems.

We do use open-source software, but all of our models are developed in-house, and we use extreme caution when it comes to training our models as well.

What do financial institutions demand of generative AI?

Mukai: Cohere works with many financial institutions. What are some of the key takeaways from those collaborations?

Baji: I think we've learned two things.

The first is that for any financial institution, information security and the protection of client information are of paramount importance. Our customers are looking to use generative AI in order to improve their productivity and client services, but that has to be achieved in an environment where information security and privacy protection are fully guaranteed.

The second is that with general-purpose AI models, it's difficult to solve the kinds of complex problems that are unique to financial institutions, which makes it necessary to provide solutions tailored to the needs of the industry and of individual financial institutions. In particular, from our collaboration with a certain influential Canadian bank customer, we've learned that there's a great need for end-to-end solutions that are aligned with the financial industry.

Mukai: Could you explain what you mean by “end-to-end solutions”?

Baji: It’s a solution that the end user can start using immediately, without any technical hassles. That means the end user doesn’t need to be concerned about connecting to a data source or uploading documents, or to worry about data or privacy issues.

In order to meet these needs, we developed a secure AI platform called “North” for companies, which can be customized for a particular industry. We developed “North for Banking” with this Canadian bank, as a solution specifically tailored to the financial industry.

North is equipped with an agentic workflow, and by connecting to a company’s various data sources, it can answer questions about the company’s own legal, financial, or HR matters, as well as questions about customer products. Using North in no way requires you to have machine learning engineers. For companies that lack AI personnel, end-to-end solutions surely have great significance.

Mukai: How do you see the future of generative AI in finance?

Baji: I firmly believe that generative AI will come to play a more important role in the financial industry. I think that in many use cases, it has the potential to generate a high ROI (return on investment). Generative AI makes it possible to obtain more accurate, rapid, and comprehensive answers, and financial institutions as well as their customers can likely reap the benefits of using it.

The ways that financial institutions use generative AI are wide-ranging, including for creating and analyzing financial statements, portfolio management, fraud detection, risk assessment and management, customer support, and more. Our customers who started investing in LLMs early on have already begun to see the effects. In 2023, a number of financial institutions launched PoCs (proof of concepts), and then in 2024, many of them went live.

Mukai: So the earlier you start, the more benefits you can expect?

Baji: I'd say getting started early allows you to accumulate knowledge more quickly and establish a competitive advantage sooner. I advise customers, "Think big, but start small". If you can rapidly solve challenges in a use case of appropriate scale, you'll gain more confidence, which will then allow you to take on more complicated use cases.

Mukai: In the past, big companies would choose a startup. But today, we live in a world where startups choose their partners. NRI is eager to explore this partnership—and we'd be thrilled to work together with you.

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 Tokyo, New York, London, Beijing and Sydney, with over 16,700 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 & Innovation 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/list.html#lakyara