NIKKEI

PRESS RELEASE

Large Language Models Specializing in Economic and Financial Information

April 24, 2024—The Nikkei Innovation Lab, a research and development department of Nikkei Inc. has developed the NIKKEI Language Model (NiLM), a series of large language models specializing in economic and financial information. Public information on the Internet was not used, only articles from about 40 years of the Nikkei, Nikkei Sangyo Shimbun, the Nikkei MJ, the Nikkei VERITAS, NIKKEI Prime, Nikkei BP and other specialized media, for which the Nikkei Group holds the copyright and usage rights. These are some of the largest language models specializing in economic and financial information in Japanese.

As of April 2024, we have finished pre-training up to 13 billion parameter models from scratch, that is, without using publicly available models as a starting point. We are evaluating the performance and the behavior by comparing them with the other models.

We have also built up to 70 billion parameter models with fine-tuning (known as continuous pre-training and instruction tuning). We have confirmed improved performance on internal tasks such as article summaries and knowledge acquisition about the latest news. Currently, the Llama 2 (70 billion) and Llama 3 (8 billion) models by Meta, Inc. are being used as the base models.

The number of tokens in the prepared Japanese corpus has reached approximately 1 trillion, without using public datasets such as Wikipedia or Common Crawl. This is one of the highest quality large language models, and one that only the Nikkei Group can provide, with its nearly 150-year history of providing economic and financial information to the world.

Large language models are beginning to be applied in many businesses, creating new opportunities. However, at the same time, several challenges have emerged. These include a lack of up-to-date knowledge, hallucinations, and the high likelihood of the data being used without permission in various media. The Nikkei continues to research and develop large

language models, using our own data, with a sense of responsibility as a news organization. Our current focus is finance and economics, creating models that can be continuously updated with the latest information.

The Nikkei Innovation Lab has continuously worked on pre-training using our own data. We keep building and exploiting pre-trained models such as RoBERTa, GPT-2, T5, and DeBERTa to keep up with technological advances. We have also made many other achievements in the field of natural language processing, including research on time-series performance degradation of language models, training data extraction, and hallucinations. For Minutes by NIKKEI, launched in the fall of 2023, we developed AI editing support tools for editors. Various achievements have been made with regard to generative AI, including the screening in Paris of an animated film produced with AI.

The developed models will be considered for use in the research and development of AI products being promoted by the Nikkei Innovation Lab. We plan to use these large language models for the domains of finance and economics, in which the Nikkei is a world leader, in various projects in the future. We will continue to improve the performance of large language models and investigate the associated challenges.

The Nikkei is continuing to discuss how to deal with generative AI as it evolves. It remains the task of human journalists to visit the scene as news unfolds, to analyze information based on their accumulated experience, and convey accurate information to customers. We will keep exploring responsible journalism, conducted by human beings, in the new era.

References

PyCon JP 2022. https://2022.pycon.jp/en/timetable/?id=EEA8FG
Shotaro Ishihara, Hiromu Takahashi, and Hono Shirai (2022). Semantic Shift Stability: Efficient Way to Detect Performance Degradation of Word Embeddings and Pre-trained Language Models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. https://aclanthology.org/2022.aacl-main.17/
Shotaro Ishihara (2023). Training Data Extraction From Pre-trained Language Models: A Survey. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. https://aclanthology.org/2023.trustnlp-1.23/

---

provider of business news and information. Founded as a market news provider in Japan in 1876, Nikkei has grown into one of the world's largest media corporations, with 37 foreign editorial bureaus and approximately 1,500 journalists worldwide. Nikkei acquired the UK-based Financial Times in 2015. Our combined digital and print circulation totals about 2.3 million, and we are continually deploying new technologies to increase our readership.

Contact
Public Relations Office
Nikkei Inc.
pr@nex.nikkei.co.jp
https://www.nikkei.co.jp/nikkeiinfo/en/