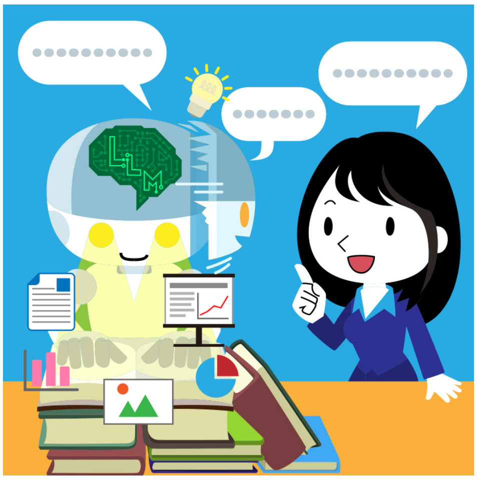
P02

It comprehends image-rich materials like manuals and brochures

“tsuzumi” for understanding graphic documents

IOWN Pick Up

NTT version Large Language Models

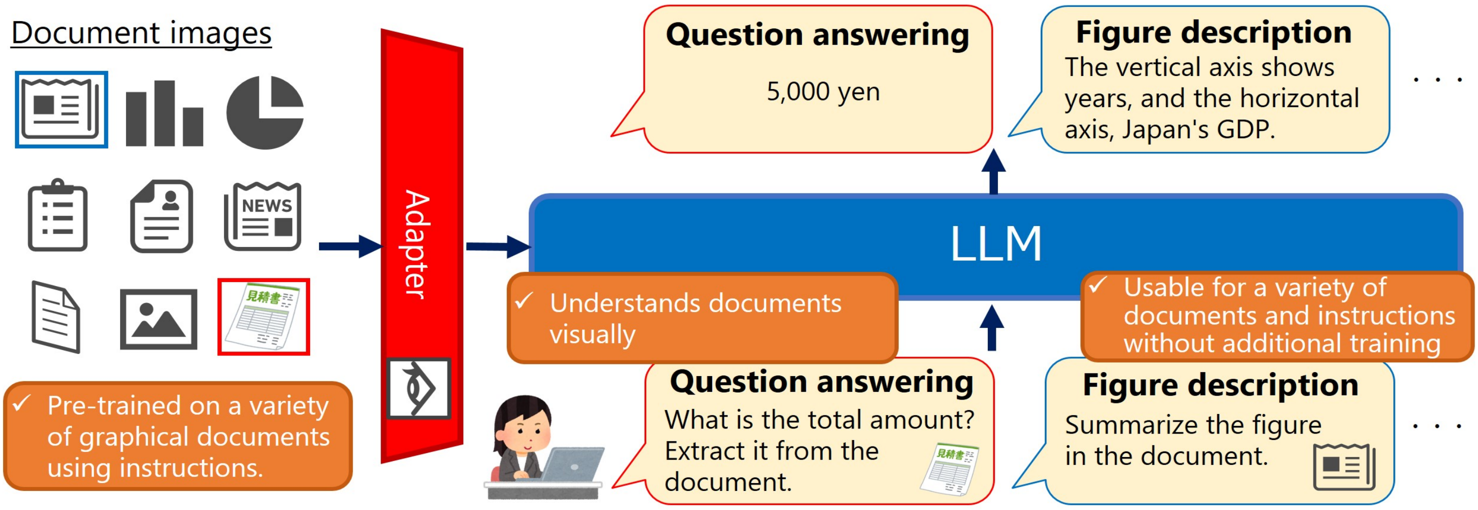


Background

Most conventional LLMs can only handle textual information and not visual information such as graphs, icons, illustrations, font size, and layout, which are usually included in documents.

Summary

Our model understands document images and presents the information sought by the user. By training it from a large number of document images, it outputs a response based on natural text instructions about the document image. It is applicable to a variety of documents and formats.



Document images

✓ Pre-trained on a variety of graphical documents using instructions.

Adapter

LLM

✓ Understands documents visually

✓ Usable for a variety of documents and instructions without additional training

Question answering
What is the total amount? Extract it from the document.

Figure description
Summarize the figure in the document.

Question answering
5,000 yen

Figure description
The vertical axis shows years, and the horizontal axis, Japan's GDP.

Features

- We lead the world in visual-document understanding and its integration with LLMs. (submitting to AAAI'24, accepted by AAAI'23 and '21, 2nd place in InfographicsVQA competition)
- Our model uses a highly accurate image encoder that we created from hundreds of millions of pairs of Japanese text and images
- We created the world's largest document image dataset and trained LLMs to understand a variety of instructions and documents, such as charts, web pages, and handwritten documents

Future_benefits

We aim to develop an AI that understands the "world as humans see it" by linking it to language and that can collaborate with humans in the office as well as other environments.

Exhibiting Company

NIPPON TELEGRAPH AND TELEPHONE CORPORATION

Contact

rdforum-exhibition@ml.ntt.com

NTT R&D FORUM 2023