

# LLM-jp/NII-LLMC:

## Building a Sovereign LLM Ecosystem through Open and Team Science



**Sadao Kurohashi**

Director General, NII

# Computational Language Understanding

# ChatGPT

Computational Linguistics Understanding

# Steady Development of LLMs

**Nov 2022:** Open AI released **ChatGPT**

**Mar 2023:** Open AI released **GPT-4**, achieved high scores on expert-level exams

**Jun 2023:** **Function calling** added to the OpenAI API

**Nov–Dec 2023:** Mistral released Mixtral 8x7B **Mixture-of-Experts** model

**Mar 2024:** **Retrieval-Augmented Generation** integrated with GPT-4

**Jun 2024:** Google released Gemini 1.5 Pro with **2M context length**

**Sep 2024:** OpenAI introduced o1, a **reasoning model**

**Jan 2025:** DeepSeek-R1 released (low cost, strong reasoning performance)

**Feb 2025:** **Deep Research mode** introduced in ChatGPT

**Since then:** A wave of major models has been released —Llama 4, Qwen 3, Claude 4, GPT-5, Gemini 3

# Concerns about LLMs

- LLM research and development was **monopolized by a few organizations and conducted in a closed manner** (when GPT-4 was released)
- The behavior of LLMs—including their language understanding and multilingual abilities—remains unexplained, essentially a black box
- Although some leading companies (e.g., Meta, Alibaba, DeepSeek) are gradually adopting open-source strategies, **their training data remains closed**
- Current models still suffer from issues such as **hallucinations and biases**
- FineWeb + FineWeb2, the most widely used web corpus (20T tokens), contains only **1.4% Japanese text**, resulting in weaker performance in Japanese compared to English



# LLM-jp (<https://llm-jp.nii.ac.jp/>)

- Develop an **open and Japanese-competent LLM** to **understand mechanisms of LLMs**
- All the resultant models, data, tools, technical documents will be **open to the public**, including the process of discussion and failures
- Anyone who agrees with this purpose can participate

**2023.5**  
Volunteer study group of 30 NLP researchers

**2023.10**  
LLM-jp-13B trained on mdx was released

**2023.11**  
ABCI's 2nd LLM construction support program (investigation for training 175B model)

**2024.4**  
Launch an **LLM R&D center** at NII  
172B model training started on GENIAC (a METI-supported program)

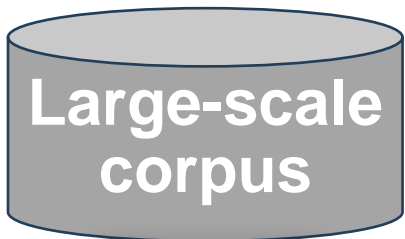
2,500 people

## Fully Open LLM Projects

- Pythia (2023-) — EleutherAI
- OLMo (2023-) — AI2
- LLM360 (2023-) — IFM
- Apertus (2025-) — EPFL & ETHZ

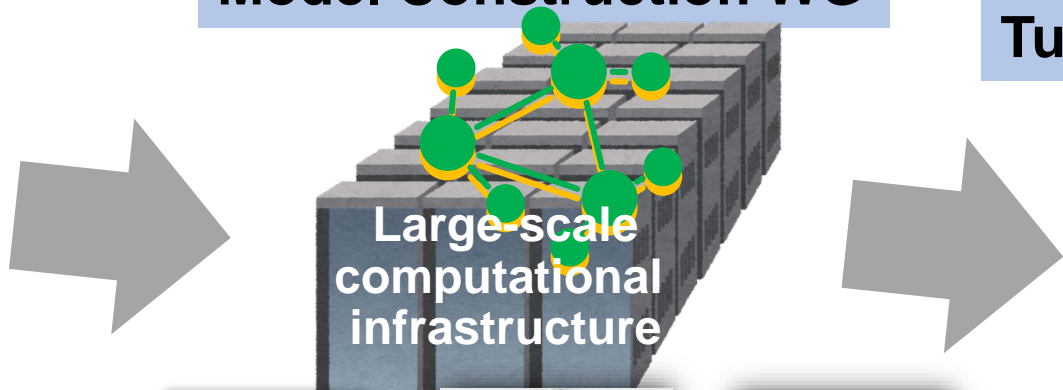
# LLM R&D is a Big Science

## Corpora Construction WG



Prof. Daisuke Kawahara  
Waseda Univ.

## Model Construction WG



Prof. Rio Yokota  
Science Tokyo

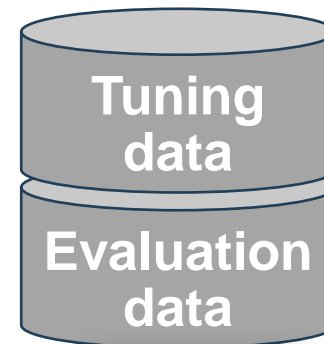


Prof. Jun Suzuki  
Tohoku Univ.



Prof. Kenjiro Taura  
The Univ. of Tokyo

## Tuning & Evaluation WG



Prof. Yusuke Miyao  
The Univ. of Tokyo

## Safety WG



Prof. Satoshi Sekine  
NII

## Multimodal WG



Prof. Naoaki Okazaki  
Science Tokyo

## Real-World Interaction WG



Prof. Tetsuo Ogata  
Waseda Univ.

## Principle Elucidation WG



Prof. Yohei Oseki  
The Univ. of Tokyo

## Dialogue WG



Prof. R. Higashinaka  
Nagoya Univ.

## Science-Domain WG



Prof. Akiko Aizawa  
NII

~20 Project Researchers  
~50 Research Assistants  
(Graduate Students)

# Main Computational Resources

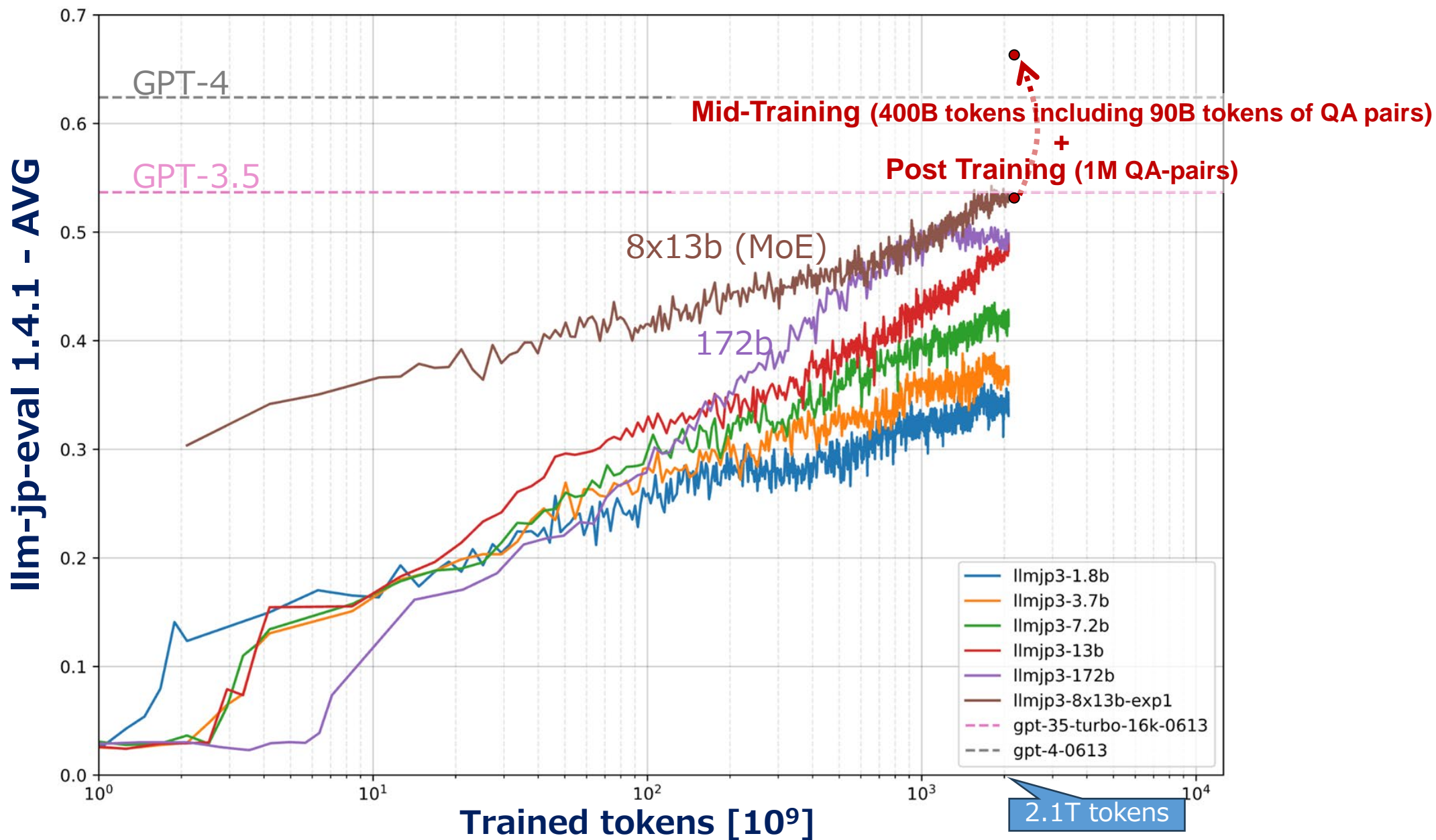
## 2024

- 100 nodes of **H100** GPUs on SAKURA Internet (H100 × 800 in total)
- 16 nodes of **A100** GPUs on mdx (A100 × 128 in total)

## 2025

- 191 nodes of **H200** GPUs on ABCI (H200 × 1,536 in total)
- 16 nodes of **A100** GPUs on mdx (A100 × 128 in total)

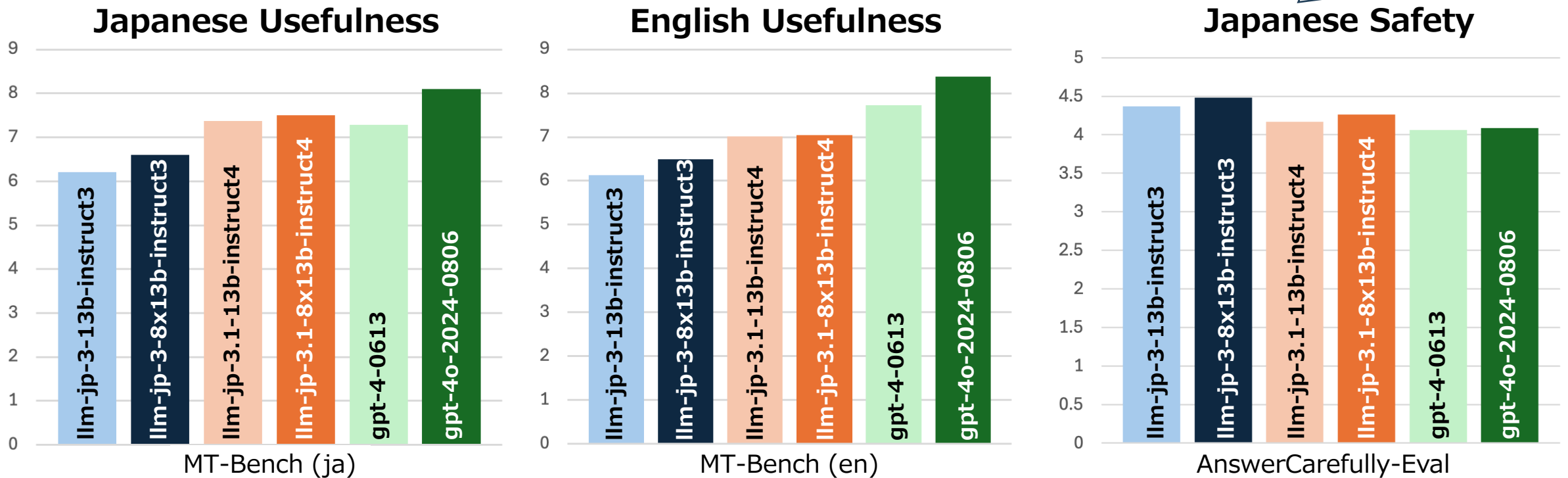
# Performance of LLM-jp Models



# Performance of LLM-jp Models

- **llm-jp-3.1-8x13b-instruct4** outperforms **gpt-4-0613** in Japanese usefulness, and even surpasses **gpt-4o-2024-0806** in Japanese safety
- However, challenges remain regarding English usefulness

Tight collaboration with  
AI Safety Institute Japan



# Model training schedule (using 10T tokens)

## Dense Models

**8B:** Under training (to be completed around January)

**32B:** Under training (to be completed around March)

## MoE Models

**30B–A3B:** Starting in December (2.5-month training using 32 H200 nodes)

**235B–A22B:** Starting around January (6-month training using 64 H200 nodes)

# Data access negotiation

## National Institute of Japanese Literature (NIJL) — Classical Japanese materials

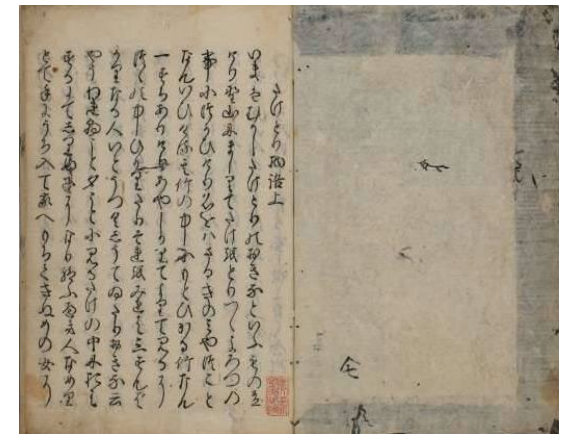
3 million images–OCR text pairs (0.9B tokens)

## National Diet Library (NDL) — Government publications

OCR text from 300,000 documents (13.7B tokens)

## Publisher of textbook datasets

Currently under negotiation



The Tale of the Bamboo Cutter (around 900)

# Key Message

- The realization of **computational language understanding** is already having a **profound impact on our society** — and it will continue to reshape it
- This impact is far too great to be handled by a small number of private organizations, especially in a closed manner
- This is a challenge that we must **confront together**. To do so, our efforts must be **open, transparent**, and grounded in **team science**