Theme: Machine Intelligence

- Sub theme: Trustability & Controllability of Large Language Model

We are seeking proposals for a research project focused on the critical area of trustability and controllability of large language models. As you may be aware, recent developments in AI, particularly large language models, have revolutionized various domains such as natural language processing, information retrieval, and human-computer interaction.

However, concerns surrounding the trustworthiness and controllability of these models have arisen. It is imperative to ensure that these powerful tools are transparent, accountable, and aligned with human values, thereby minimizing potential risks and fostering responsible AI deployment.

The objective of this research project is to investigate and address the following key areas :

- Trustability of Large Language Models:
  a. Assessing the reliability and accuracy of generated outputs.
  b. Developing techniques to identify and mitigate bias, misinformation, and harmful content.
  c. Exploring methods to increase the interpretability and explainability of model behavior.
- Controllability of Large Language Models:
  a. Investigating mechanisms for user control and intervention during model operation.

b. Studying techniques to enable customization of model behavior to align with specific requirements and ethical guidelines.

c. Examining methods to restrict undesirable outputs while maintaining model performance.

※ The participants are also encouraged to propose new ideas outside the topics listed above.
※ Funding: Up to USD 150,000 per year