Towards Building Multicultural and Multilingual Safe Large Language Models

On Monday, November 11, 2024, we held a debriefing session on our participation in the AI Safety Red Teaming Challenge Project organized by Singapore's Infocomm Media Development Authority (IMDA). This webinar aimed to share the results and challenges of Red Teaming efforts to assess and improve the safety of AI models, addressing region-specific risks and concerns related to generative AI.

Overview

- Date: Monday, November 11, 2024, 10:00–11:00
- Venue: Zoom Webinar
- Organized by: Tokyo College, The University of Tokyo; Institute for Future Initiatives, The University of Tokyo; Next Generation Artificial Intelligence Research Center, The University of Tokyo; B'AI Global Forum, The University of Tokyo; Japan Deep Learning Association
- Supported by: National Institute of Informatics, Research and Development Center for Large Language Models; Japan AI Safety Institute

Background

As generative AI becomes more prevalent, it is vital for AI models to reflect region-specific cultural and linguistic risks while ensuring safety. These evaluations often fail to account for the diverse cultural and linguistic contexts in regions outside the West, creating a gap in addressing risks effectively. The Singapore government-led AI Safety Red Teaming Challenge Project aims to enhance AI safety by integrating diverse viewpoints from participating countries, including Japan.

Summary

The webinar began with opening remarks by Arisa Ema (Tokyo College, The University of Tokyo), who emphasized the importance of addressing risks in LLMs, such as harmful content and cultural biases. She highlighted the need for a Japanese perspective in tackling these challenges and fostering collaboration among stakeholders to enhance generative AI safety and governance.

The webinar featured two key presentations:

- Satoshi Sekine (NII-LLMC/RIKEN AIP), who led the Japanese delegation to the AI Safety Red Teaming Challenge Project, provided insights into the progress and outcomes of efforts to advance AI model safety.
- Teresa Tsukiji (Japan Deep Learning Association) detailed Japan’s participation in the AI Safety Red Teaming Challenge Project, highlighting key contributions and outlining future directions for improving AI safety.

In the latter half of the event, an interactive Q&A session provided an opportunity for participants to engage in dynamic discussions. Key topics included strategies for mitigating risks associated with generative AI and practical approaches to addressing region-specific challenges. These discussions underscored the importance of striking a balance between AI safety and usability, while fostering collaborative efforts across diverse communities.

実害にはどのようなものがあるか
ョン、政策などに関する研究者や実務家
ガバナンスに関心を持つ方々の
を形成

From right: Arisa Ema, Satoshi Sekine, Teresa Tsukiji