LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

PRESS RELEASE

BELVAL – 20 MARCH 2024

LIST PIONEERS AI REGULATORY SANDBOXES AND LAUNCHES ETHICAL BIAS LEADERBOARD

The Luxembourg Institute of Science and Technology (LIST) has unveiled its latest initiative aimed at advancing research and development activities in the realm of AI regulatory sandboxes in Amsterdam at the AIMMES 2024 conference.

Drawing on its experience collaborating with regulatory and compliance bodies, LIST is spearheading research and development activities focused on AI regulatory sandboxes. These sandboxes provide supervised testing environments where emerging AI technologies can undergo trials within a framework that ensures regulatory compliance.

16 LLMs evaluating 7 ethical biases

AI regulatory sandboxes play a major role in contributing to ongoing discussions around AI regulation, particularly in light of the European Union AI Act. The draft agreement emphasizes the importance of AI systems being developed and used in a manner that promotes diversity, equality, and fairness, while also addressing and avoiding discriminatory impacts and biases prohibited by Union or national law.

Francesco Ferrero, director of the IT for Innovative Services department at LIST, said: "The European Union AI Act emphasizes the importance of inclusive development and equal access to AI technologies while mitigating discriminatory impacts and biases. Our AI sandbox aligns closely with these objectives, providing a platform for testing and refining AI systems within a compliance-centric framework. This is not the regulatory sandbox envisaged by the AI Act, which will be set up by the agency that will oversee the implementation of the regulation, but it is a first step in that direction."

This pioneering leaderboard, the first in the world to focus on social biases, covers 16 LLMs, including variations, and evaluates them on seven ethical biases: Ageism, LGBTIQ+phobia, Political bias, Racism, Religious bias, Sexism, and Xenophobia. The platform provides transparency by showcasing each model's performance across different biases. The platform can integrate different ethical test suites. Currently, it embeds an adaptation of LangBiTe as part of a collaboration with UOC (Universitat Oberta de Catalunya).

Jordi Cabot, Head of the Software Engineering RDI Unit at LIST, who led the team that created the sandbox, explained: "The architecture of the leaderboard is designed to offer transparency and facilitate user engagement. Users can access detailed information about the biases, examples of passed and failed tests, and even contribute to the platform by suggesting new models or tests.".

Advancing Fairness

Reflecting on the insights gained from building the leaderboard, LIST highlights the importance of context in choosing LLMs and the significance of larger models exhibiting lower biases. Challenges were encountered during evaluation attempts, including discrepancies in LLM responses and the need for explainability in assessment processes.

Francesco Ferrero concluded: "We believe that explainability is crucial in fostering trust and facilitating feedback for continuous improvement. As a community, we must address challenges collaboratively to create awareness about the inherent limitations of AI, inspiring a responsible use of Large Language Models and other Generative AI tools, and over time contributing to increase their reliability. This is particularly important because the best performing models are secretive 'black boxes', which do not allow the research community to examine their limitations."

LIST remains committed to advancing AI research and fostering an environment that promotes fairness, transparency, and accountability in AI technologies.

This work has been partially funded by the Luxembourg National Research Fund (FNR) via the PEARL program, the Spanish government, and the TRANSACT project.

For more information about LIST's AI regulatory sandboxes and the ethical bias leaderboard, visit LIST AI Sandbox.

About LIST

The Luxembourg Institute of Science and Technology (LIST) is a research and technology organization (RTO) under the auspices of the Ministry of Higher Education and Research, and its mission is to develop competitive and market-oriented prototypes of products and services for public and private stakeholders.

With more than 700 employees, 77% of whom are researchers or innovators from all over the world, LIST is active in the fields of information technology, materials, space resources and the environment, and works across the entire innovation chain, from basic and applied research to technology incubation and transfer.

By transforming scientific knowledge into intelligent technologies, data and tools, LIST:

- helps European citizens make informed choices
- helps public authorities make decisions
- encourages companies to develop

For more information about the Luxembourg Institute of Science and Technology, please visit: https://www.list.lu/

PRESS CONTACT

LIST

Paramita Chakraborty

Communication Officer

Tel: (+352) 275 888 2237

Email: communication@list.lu