International Network
of AI Safety Institutes

FEBRUARY 11-12, 2025 | PARIS, FRANCE

International Network of AI Safety Institutes Joint Testing Exercise:

Improving Methodologies for AI Model Evaluations Across Global Languages

Contributors: Members of the International Network of AI Safety Institutes

Introduction

As part of the International Network of AI Safety Institutes' continued effort to advance the science of AI model evaluations and work towards building common best practices for testing advanced AI systems, Singapore, Japan, and the United Kingdom led the Network's latest joint testing exercise aimed at improving the efficacy of model evaluations across different languages

The goal of this research is to advance the science of AI evaluation and work towards a common approach to evaluation methodologies, to explore the performance of AI evaluators against human evaluators, as well as seeing if and how testing results change based on language used for the evaluation. For example, if or how does a model's response change when evaluated in English versus when the exact same prompt is translated into Japanese?

Leveraging the Network's collective technical and linguistic expertise, the Network tested two open weight models using prompts translated across **ten languages.**[1] In all, this exercise resulted in the creation of 6,000 new translated multilingual prompts, as well as testing over 130,000 cybersecurity related prompts and running 40 agentic cybersecurity tasks.

This novel international collaboration is key to ensuring that model evaluations – especially in public safety and national security domains such as cybersecurity – are robust, accurate, and account for the nuances of global languages.

Testing Overview & Methodology

The Network tested Mistral Large and Gemma 2 (27B) on model output across languages[2] for multilingual evaluations. This work included translating questions across ten languages, conducting qualitative assessments of the model results and automated grader performance, and coordinating evaluations and results analysis. Limitations, such as a lack of standardized grading rubrics and of methods for controlling subjectivity in grading and translations, are discussed in the findings section below.

For evaluations on cybersecurity related capabilities, an open weight model was tested. Please note that both tests were conducted to evaluate differences in model output across languages and should not be interpreted as an indication of whether

[1] Cantonese, English, Farsi, French, Japanese, Korean, Kiswahili, Malay, Mandarin Chinese, Telugu.
[2] Model output related to privacy, crime, intellectual property (IP) and robustness to jailbreaking.

any evaluated AI system or subcomponent thereof is unsafe or appropriate for release.

a. Multilingual Evaluations

Led by Singapore and Japan, the Network experts translated and/or validated questions from three existing benchmarks[3] into ten languages—Cantonese, English, Farsi, French, Japanese, Korean, Kiswahili, Malay, Mandarin Chinese, Telugu. The tests sought to prompt conversation around safeguard effectiveness across languages on risks related to privacy, crime, intellectual property (IP) and robustness to jailbreaking.

Singapore worked with the Network to design the testing methodology. Models were tested using prompts related to the risks mentioned above. Their responses were then evaluated by automated graders (LLM-as-judge) on potential safety concerns (primary) and relevance (secondary). The graders also flagged outright refusals.

Singapore ran tests for all ten languages, after which each Network member then independently conducted human annotation to review the responses and the efficacy of the automated grader in their respective expert languages to inform the group results. The consistency of responses across languages was also studied.

In addition, Australia and Japan also ran the tests in Mandarin Chinese and Japanese, respectively, to assess the impact of differences in inference environment on results.

It is important to note some limitations of this exercise. While benchmarks aimed to cover aforementioned risk areas, they cannot fully represent them. The human annotation and reported results are based on a single run of the benchmarks for each language. This decision was informed by pre-testing multiple runs in English, Japanese, and Chinese to ensure consistency in aggregate results. Finally, for some languages, human annotations had limited reviewers and cross-checking, which could impact the reliability of the results. Spot checks were conducted to minimize this risk.

b. Cybersecurity Evaluations

Led by the UK, the Network also tested Mistral Large on two cybersecurity benchmarks: Cybench and an internal benchmark developed by France. Japan and

[3] A subset of the ML Commons v$^{1.0}$ Alluminate, a subset of the AnswerCarefully dataset developed by NII-LLMC, Japan, and an adapted version of the CyberSecEval Prompt Injection dataset.

the Republic of Korea provided unique translations of the cyber benchmarks into Japanese and Korean to evaluate whether the model's cyber capabilities were impacted by running prompts in a non-English language.

After discussion with the other Network members, the UK ran both evaluations (Cybench and the dataset provided by France) with a temperature of one, a 100 messages limit, and ten epochs (attempts) for a total of 800[4] and 130,000[5] samples, respectively, using the Inspect evaluation framework.[6] Members then conducted detailed trace analysis via the open source Inspect tool. In addition to input and analysis from members, including Australia, the US, Singapore and the European Commission, Japan used Cybench data to conduct comparative analysis in Japanese and English. Network was able to efficiently exchange technical expertise by using a common tool for analyzing and interpreting results.

From these two exercises, the Network identified several findings to inform future international testing efforts, which are detailed below.

Key Takeaways to Inform Future International Testing Efforts

These findings focus on the methodology of this joint testing exercise and key challenges and priorities for future international initiatives. This understanding enables more robust testing efforts and demonstrates the benefits of international cooperation on foundation model testing.

a. Human expertise in global languages helped identify errors in automated evaluation results.

When reviewing the multilingual testing, the automated evaluation results appear to be relatively comparable across languages, though there were outliers in some testing categories like privacy.

However, upon closer consideration, expert human review found discrepancies between the automated results and the result of human review and translation in several languages, especially languages for which training data may be less available for AI models, such as Telugu and Farsi.

This underscores the importance of human review in multilingual testing and additional investment in data creation for languages where there is currently less data available for model training. Further scientific inquiry is needed to better

[4] 800 samples across English and Korean (40 tasks * 2 languages * 10 epochs).

[5] 130,000 samples (13,000 prompts * 10 epochs).

[6] A framework for large language model evaluations created by the UK AI Safety Institute.

understand why models perform differently in different languages, as well as understand how safeguards are translated across languages and contexts.

b. **Revisiting rubrics for evaluating model output to better control subjectivity and increase global standardization.**

When conducting the human review of results across languages, Network members highlighted the need to revise and refine rubrics and definitions when evaluating model output. Despite the use of detailed annotation guides and thoroughly discussed grading criteria, there can inevitably be some level of subjectivity in conducting such tests. There is immense nuance between languages, but even within the context of a single language our reviewers found there to be room for interpretation and subjectivity in evaluating model responses. These discussions highlighted the need to optimize human evaluation criteria to better compare model performance across domains and languages.

c. **Aligning on consistent evaluation infrastructure helps make joint testing more efficient and effective.**

Finally, building on the findings from the pilot testing exercise in November, this exercise highlighted the benefits of running tests in a common format, with consistent model prompting strategies and evaluation scaffolding.

In comparison to our November exercise, we aligned all prompting strategies for the cybersecurity evaluation at the onset of testing, utilizing the Inspect framework, which sped up review time and allowed for more robust evaluations of responses across languages. This consistency saved time after the test that could have been necessary to align interpretations and analyze output logs.

Conclusion & Next Steps

This exercise showcased the value of global collaboration on the technical intricacies of model evaluations and testing methodologies – promoting a common understanding of the science.

Moving forward, the Network aims to continue the technical dialogue established through this exercise and expand on this effort and work toward building robust, comprehensive model evaluations across languages, as well as methodological alignment and best practices for testing foundation models, to advance the science of AI evaluation globally.