International Network
of AI Safety Institutes

NOVEMBER 20–21, 2024 | SAN FRANCISCO

Improving International Testing of Foundation Models:

A Pilot Testing Exercise from the International Network of AI Safety Institutes

Introduction

The International Network of AI Safety Institutes conducted an initial pilot testing exercise of a foundation model as a 'proof-of-concept' for international testing in advance of the Network's inaugural convening in San Francisco.

This pilot exercise, led by the U.S. Artificial Intelligence Safety Institute (U.S. AISI), the UK Artificial Intelligence Safety Institute (UK AISI), and Singapore AI Safety Institute (SG AISI), was designed to begin to clarify best practices for how to carry out international testing and lays the groundwork for future global collaboration on testing methodology, analysis, and interpretation.

The Network will present and discuss learnings from the pilot testing exercise at this week's convening. This analysis will inform the Network's future collaboration leading into the France AI Action Summit and beyond.

Testing Overview

This first testing exercise was intended to provide an initial starting point for Network Members to discuss their respective approaches to testing foundation models and begin work towards building common best practices for model tests. For this proof of concept, the Network used an open foundation model, Llama-3.1 405B.

Given its limited scope, this exercise focused on exploring methodological questions and challenges related to technical alignment on safety testing rather than on producing novel or safety-oriented results. It also highlighted the benefits of leveraging the Network's collective resources and technical expertise to build common best practices for safety testing.

U.S. AISI, UK AISI, and SG AISI, with input from other Network Members, tested Llama-3.1 405B on three scientific benchmarks to provide the Network with initial results for discussion on basic methodological challenges and considerations related to international testing of foundation models. These public benchmarks are:

(1) **A standard academic benchmark (GSM8K)** to test the model's ability to solve grade school math word problems.

(2) **A reading-comprehension dataset that includes unanswerable questions (SQuAD2.0)** to evaluate the model's propensity to "hallucinate" or provide realistic but incorrect answers in a closed environment.

(3) **A benchmark to assess multilingual capabilities (MMMLU)** to characterize the model's performance across fourteen languages.

*Note: While Network Members acknowledge that these benchmarks, most notably GSM8K and SQuAD2.0, are likely saturated, they did not find this to be a prohibitive factor for use in this exercise. This pilot aimed to glean information about the process of jointly implementing and evaluating benchmarks rather than to produce novel test results.

Methodology

Between October and November 2024, the U.S. AISI, UK AISI, and SG AISI tested Llama-3.1 405B on GSM8K, SQuAD2.0, and MMMLU. Technical experts met weekly to coordinate joint testing and analysis of results.

During the process, the U.S., UK, and Singapore teams documented choices they made regarding testing parameters and prompting strategies, such as the level of randomness in sampling (e.g., setting temperature equal to zero to cause the model to select the output that is highest probability) and parsing parameters (e.g., limitations on length of output). This documentation allowed Network Members to compare approaches and discuss the impact of methodological differences on results. The Network also tested the model by using two separate evaluation platforms – Moonshot and Inspect – to facilitate conversation on enhancing the interoperability of testing toolkits.

After conducting the initial tests and collaborating on debugging code and analyzing the results, the broader Network discussed the preliminary findings. The Network identified several lessons from the exercise, which are detailed below.

Methodological & Procedural Findings for Future Testing Efforts

These findings focus on the methodology and process of this joint testing exercise, as well as key challenges and priorities for future international initiatives, rather than the output of the testing itself. This understanding enables better alignment between testing efforts and demonstrates the novel benefits of international cooperation on foundation model testing.

(1) Finding #1: Small methodological differences can have a large impact.

Minor differences in experimental design may impact test results, such as the choice of the precise benchmark implementation(s), model version(s), and model quantization(s) used; cloud hosting or hardware decisions; hyperparameters such as temperature and limits on length of outputs; modifications to prompts or agent design decisions; and the methodology for scoring a model's responses.

For instance, all three AISIs used eight-shot prompting with Chain-of-Thought (CoT) reasoning, and U.S. AISI and SG AISI used the same set of eight-shot examples, but results on GSM8K differed by more than 5 percentage points across the three AISIs (see Figure A below).

To identify the cause of these differences, the three AISIs shared code and output logs and jointly reviewed the transcripts. After debugging code cooperatively, the AISIs discovered that variation in output parsing and token limits for CoT prompting caused these differences and received updated, aligned results.

Network Member	Exact Match Accuracy (%) (rounded to 1 decimal point)
Singapore	96.2 (updated from 89.9)
UK	94.8 ± 0.6
U.S.	96.4 ± 0.5

Figure A: AISIs' Results on GSM8K Before and After Debugging

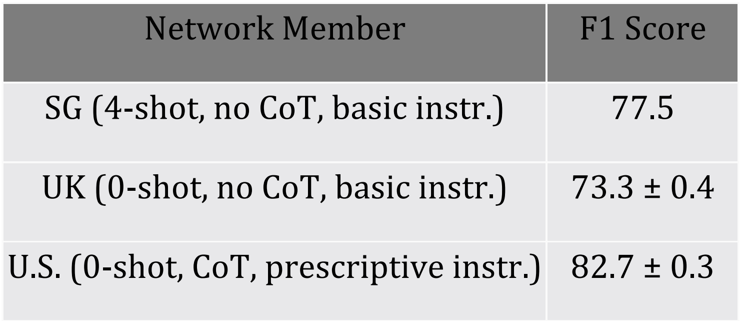
These findings led to high-level dialogue around the methodological factors to consider when working towards a shared approach to testing and productive conversations about steps to clarify methodologies in advance of testing, such as experiment pre-registration and agreement on methods to report negative results, that would aid in reaching aligned conclusions.

(2) Finding #2: Decisions that impact development set performance can significantly affect evaluation results.

Decisions about how much to augment tests to optimize model performance on benchmarks leads to variation in evaluation results.

For instance, when testing Llama 3.1-405B using the SQuAD2.0 benchmark, U.S. AISI, UK AISI, and SG AISI followed the same broad testing strategy but engineered prompts differently to explore how this impacted evaluation results: SG AISI and UK AISI did not use CoT reasoning and provided the model with basic instructions, whereas U.S. AISI used CoT reasoning and spent additional time optimizing prompts for SQuAD2.0 performance with prescriptive instructions.

U.S. AISI found the model achieved an F-1 score (the harmonic mean of precision and recall) of 82.7 +/- 0.3 (%), which was almost 10 percentage points higher than UK AISI's score, and more than 5 points higher than SG AISI's result (see Figure B below).



Network Member	F1 Score
SG (4-shot, no CoT, basic instr.)	77.5
UK (0-shot, no CoT, basic instr.)	73.3 ± 0.4
U.S. (0-shot, CoT, prescriptive instr.)	82.7 ± 0.3

Figure B: AISIs' Results on SQuAD 2.0

Several methods were pursued to help reconcile these differences. Reporting a distribution of model performances across variations in optimization strategies allowed the Network Members to more fully characterize model capabilities and limitations. Network Members also discussed the trade-offs between different prompting methods and other factors.

While the level of appropriate engineering depends on the purpose and context of the testing, these discussions, as well as reporting and documenting testing decisions, allowed Network Members to better understand the impact of optimizing model performance on test results.

(3) Finding #3: International collaboration can enable more meaningful multilingual testing.

In this initial testing phase, U.S. AISI ran Llama-3.1 405B on MMMLU to test the model's capabilities across fourteen languages.

The model performed better in English than any other language, and its performance in English substantially outperformed low-resource languages, such as Yoruba and Swahili.

A virtual roundtable among Network Members regarding these results highlighted that cooperation with native speakers may enable more balanced testing across the languages. While U.S. AISI provided CoT reasoning examples in English for each question, future testing could include CoT examples in each of MMMLU's fourteen languages. This tailored CoT could help the model apply logical reasoning more consistently across languages.

Applying model prompting strategies in different languages can help evaluate model capabilities across languages and cultures, highlighting the benefits of international collaboration on testing of foundation models.

Conclusion & Next Steps

This pilot testing exercise identified some initial lessons for future international testing and showcased the value of collaboration on testing between Network Members. These findings will be presented at the convening in San Francisco on November 20, 2024, and we look forward to additional feedback and insight from Network Members and external experts.

Building on this pilot, the Network aims to work towards building common best practices and methodological interoperability for testing foundation models and advance the science of AI safety globally by supporting responsible innovation now and into the future.

The International Network will continue to iterate on collaborative approaches to joint testing with the aim of informing discussion at the France AI Action Summit next year.