



[AI-PHI] 23th SESSION

AI News

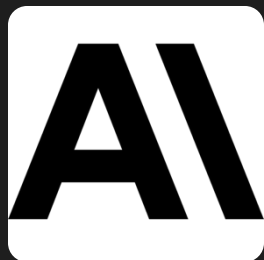
AI-Phi Community

29/04/2025

ANNOUNCEMENTS

- **Causerie Output Ready**
- **Design overhaul - sometime later this year**
- **Next session by Ryota Kanai**
 - **Artificial Consciousness**

The Urgency of Interpretability

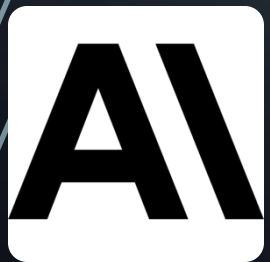


<https://www.darioamodei.com/post/the-urgency-of-interpretability>



Dario Amodei

“The Urgency of Interpretability”



Dario Amodei

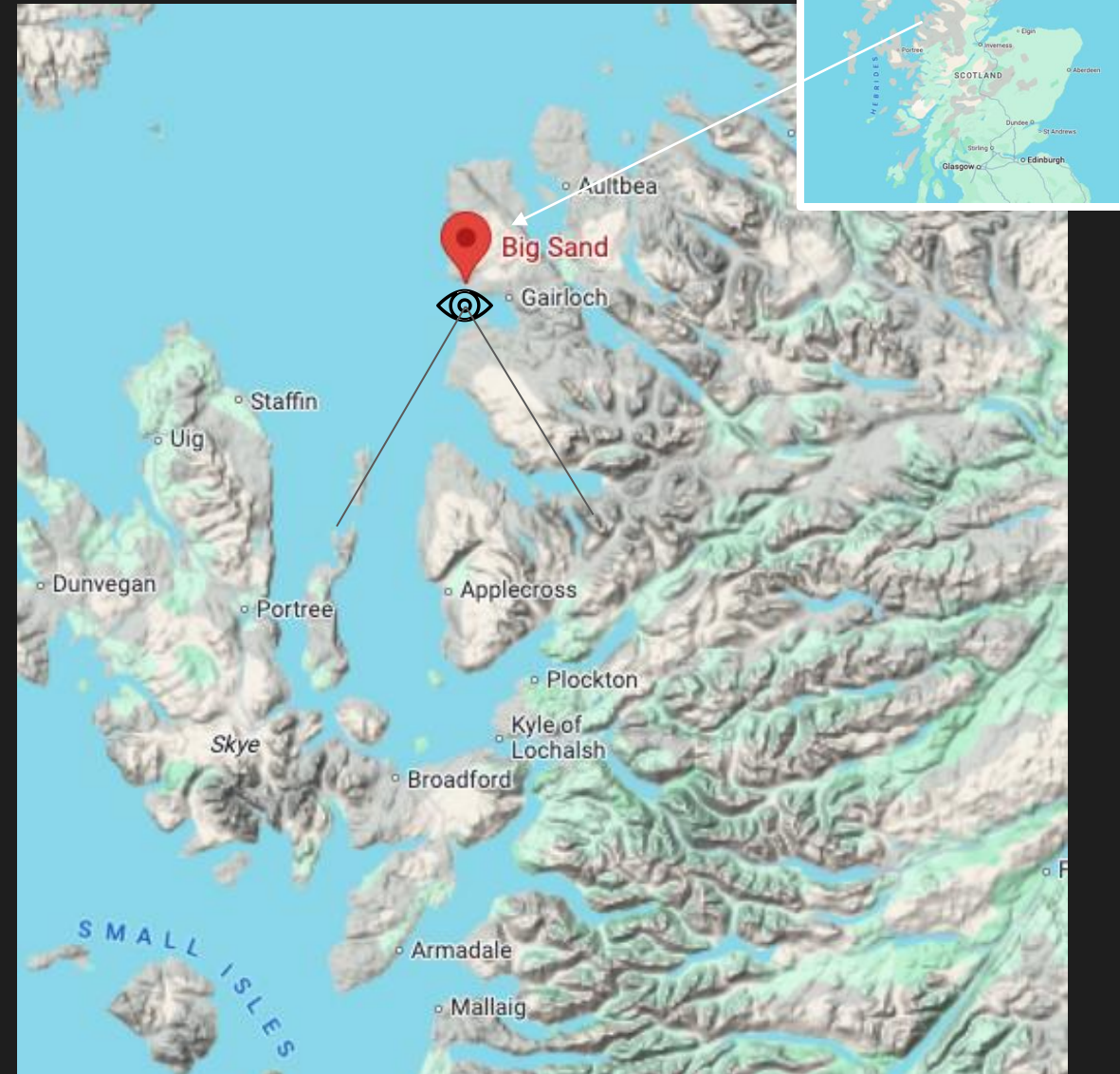


I'm playing a geoguesser game. Can you tell me what the geo location of the following location is?

Based on the visual clues:

1. **Landscape:** Wide sandy beach with prominent sand ripples (indicating strong tides/winds), a body of water (sea loch), and a distinctive range of rugged mountains in the background, some with snow patches.
2. **Mountains:** The jagged, rocky peaks are characteristic of the Torridon Hills or similar ranges in the Northwest Highlands of Scotland.
3. **Location:** This specific view, with the large expanse of sand and the particular profile of the mountains across the water, is very recognizable.

This is **Big Sand beach (Tràigh Mhòr) near Gairloch, Wester Ross, Scotland**. The mountains visible across Loch Gairloch are part of the Torridon range.



✦ Thoughts

Auto

The thoughts produced by the model are experimental

1. **Analyze the Image:** The image shows a wide sandy beach with distinct ripple patterns, likely formed by tides. In the background, there's a body of water (likely the sea or a large loch) and a prominent mountain range. The sky is partly cloudy but mostly blue. There are a few distant figures on the beach. The vegetation on the low hills to the left looks like typical coastal Scottish/UK moorland. The mountains have some snow patches on the peaks, suggesting a cooler climate or higher altitude.
2. **Identify Key Features:**
 - Wide, rippled sandy beach.
 - Distinctive, rugged mountain range in the background.

AI-Scientist-v2 – Passes Peer Review at ICLR

International
Conference on
Learning
Representations

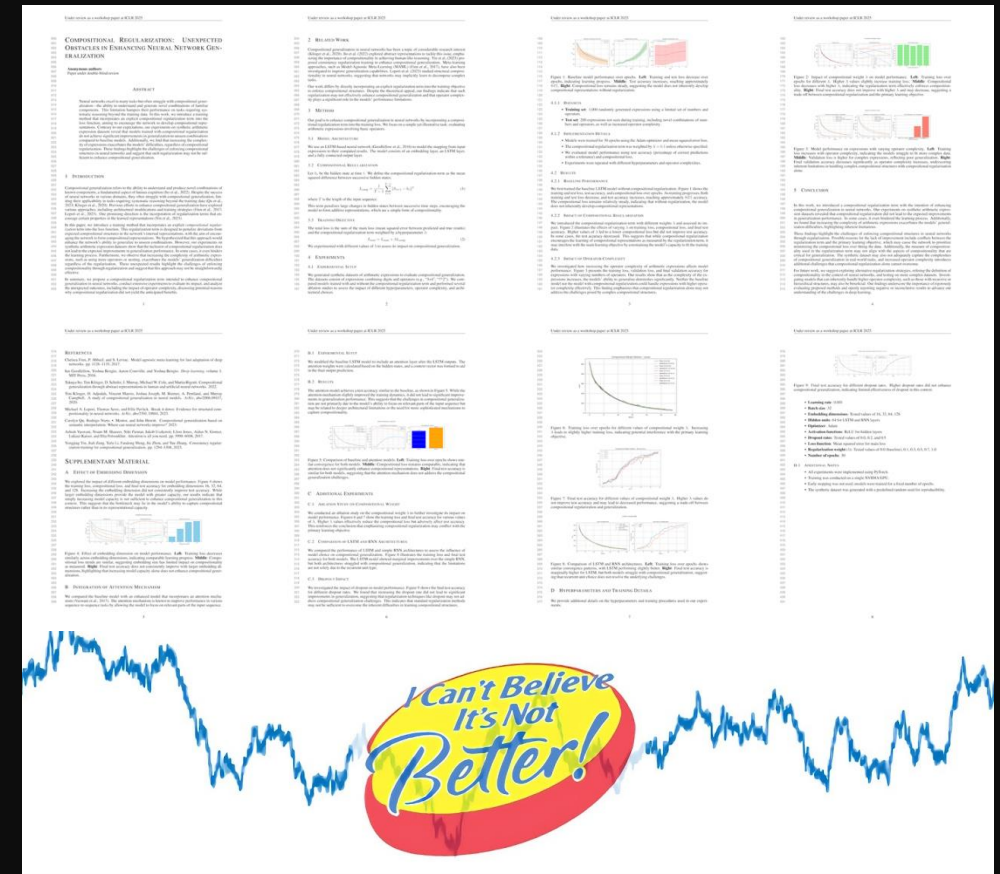
Reviewers knew there might be AI content

3 'Best' papers submitted. One accepted.

Input: Broad topic of the workshop

Output: Entire Paper

Hypothesis, experiments, running experiments,
analysis, formatting, figure creation and placement.



Rating: 6: Marginally above acceptance threshold

Rating: 7: Good paper, accept

