

✱ Apart Research

A research lab and advocacy
organization for AI safety

Enabling, researching, and deploying AI safety

Apart Research

1

AI safety research & fellowship

We incubate research talent through the accredited and structured academic publication pipeline

2

Research hackathons

We host AI safety research hackathons virtually and across seven continents for mid-career scholars

3

Exploratory AI safety work

We collaborate with organizations and institutions to explore new avenues to reduce AI risks

We unlock potential and advance AI safety

- (1) Speaking at conferences [1, 2]
- (2) Sharing researcher's work and keynote on YouTube [@apartresearch](#)
- (3) Mentoring AI safety with [Condor Camp](#), [EntrepreneurFirst](#) and [Meta](#) [1]
- (4) AI Safety Ideas ([aisi.ai](#))
- (5) [Reading What We Can](#)
- (6) [Interpretability starter](#)
- (7) [Safety Timelines research](#)
- (8) [AI cyberdefense](#)
- (9) [discontinued] Weekly newsletter in AI safety
- (10) [discontinued] Weekly YouTube update videos in AI safety

[apartresearch.com](#)

contact@apartresearch.com to connect with us

* Apart Lab

We conduct technical research in AI safety. We help talented entrants to the field of AI safety become lead authors to provide experience with academic research. These often come from the hackathons.

We have 4 papers under review, 2 in progress, and 3 already published. See the publications here:

(1) **Neuron to Graph: Interpreting Language Model Neurons at Scale**

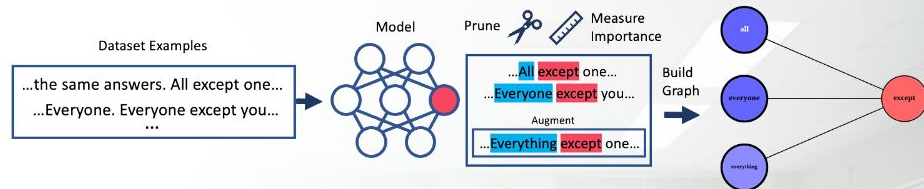
Alex Foote*, Neel Nanda, **Esben Kran**, Ionnis Konstas, Shay Cohen, **Fazl Barez*** published at the RTML workshop at ICLR 2023

(2) **Detecting Edit Failures In Large Language Models: An Improved Specificity Benchmark**

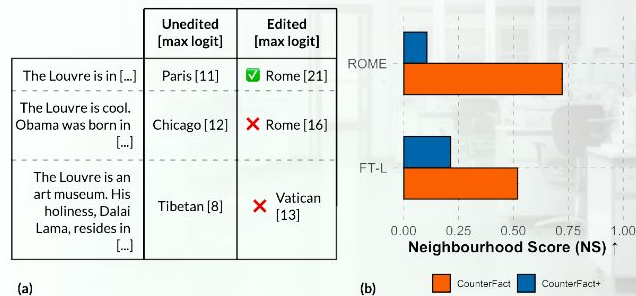
Jason Hoelscher-Obermaier*, **Julia Persson***, **Esben Kran**, Ionnis Konstas, **Fazl Barez*** published at Findings of ACL 2023

(3) **The Larger They Are, the Harder They Fail: Language Models do not Recognize Identifier Swaps in Python**

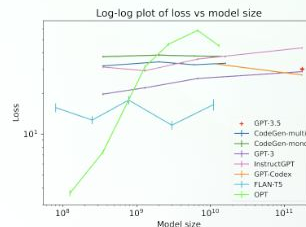
Antonio Valerio Miceli-Barone*, **Fazl Barez***, Ioannis Konstas, Shay B. Cohen published at Findings of ACL 2023



[1] Foote et al., 2023



[2] Hoelscher-Obermaier et al., 2023



"Doing research with Apart Lab has been a great experience! I found it extremely helpful to have a forum for collaboration and discussion. The interactions with my collaborators at Apart Lab helped me stay motivated and see the project through to its successful publication. The project has strongly increased my confidence that I can contribute to AI safety research. I highly recommend working with Apart Lab!"

- Hoelscher-Obermaier

[3] Miceli-Barone et al., 2023

What we do

Alignment Jam hackathons

- 3-day output-focused workshops in AI safety
- Hosted online and at in-person sites concurrently
- We organize keynote speakers, judges, live streams, preparatory reading group resources, starter code notebooks, announcements, and more while local groups organize their space
- Organize a hackathon with us by connecting at jam@apartresearch.com

From participants

9.15% increase in AIS career probability after participating
8.7 / 10 would recommend joining to a friend
1,050 members in our community server

168 projects submitted by over 400 people across 10 Alignment Jams and 3 workshops hosted at 69 events in collaboration with AI safety groups in more than 30 in-person locations

alignmentjam.com

contact@apartresearch.com to connect with us



With collaborators from



Apart Research

Connect at contact@apartresearch.com



Esben Kran

Co-director (CEO)

3x founder. Product (lead data scientist),
Interacting Minds Centre, Center for Humanities
Computing, EKC Design. Board of EA Denmark.



Fazl Barez

Co-director (CTO)

PhD (Edinburgh / Oxford). FLI Fellow, Turing
Fellow. Amazon, Huawei Research, AI Safety
Hub, CSER, Data Lab. Advisor to EA Edinburgh.

Scholar researchers

The fellows of our Apart Lab program are treated as active researchers within the lab and so receive opportunities for talks, collaborations, and more

Alumni researchers

We have multiple alumni who are still helping out with research and mentoring the next generation of scholars

Hackathon site volunteers

Local AI safety groups help us out on their own time – see alignmentjam.com/map

Other volunteers

A few people help us with maintaining our websites along with Spanish translations

Community

Our Discord community helps each other out a lot

* Apart Research

How can we help you

Co-hosting hackathons

We can partner to host your research workshops in AI safety with our [hackathon platform](#), our network of keynote speakers and our 1,000+ member community of technical students, researchers, and engineers. Previous partners include:

- Condor Camp
- AISafety.info
- Agency Foundations

Contact us at hack@apartresearch.com

Consulting & speaking arrangements

With our experience in machine learning safety, mechanistic interpretability, data science, and benchmarking and monitoring safety, we can help your organization with AI safety research and impact-focused product R&D. Contact us for speaking events, mentoring, or hosting a workshop for your organization. Previous partners include:

- EAGx Cambridge, Nordics, and Warsaw Conferences
- Foresight Institute
- Meta and EntrepreneurFirst
- Condor Camp
- Tality

Contact us at consulting@apartresearch.com

Apart Research

- Impact overview (apartresearch.com/impact)
- Alignment Jam (alignmentjam.com)
- Apart Research (apartresearch.com)