

# IMPERIAL

## Research Associate in Detecting Unverbalized Dangerous Capabilities in Frontier Models

|                              |                                                                                                                                |
|------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Department/Division/Faculty: | Department of Computing, Faculty of Engineering                                                                                |
| Campus/Location:             | South Kensington, London                                                                                                       |
| Job Family/Level:            | Research Job Family, Research Associate ( <a href="#">Salary Scale</a> )                                                       |
| Responsible to:              | Dr Oana-Maria Camburu                                                                                                          |
| Line management for:         | N/A                                                                                                                            |
| Key working Relationships:   | Internal staff and students within the Department of Computing and industry partners                                           |
| Contract type:               | Full-time, Fixed-term 1-year to start as soon as possible and before 8 <sup>th</sup> April 2027 with possibility of extension. |

### Purpose of the Post

We are looking for motivated Postdoctoral Research Associates (PDRAs/PostDocs) to join the [Safety&Alignment](#) lab at Imperial College London, led by **Dr. Oana-Maria Camburu** in a fully funded research role.

**Overview.** The position is fully funded in the context of Dr. Oana-Maria Camburu's grant on the project **Detecting Unverbalized Dangerous Capabilities in Frontier Models**. The project focuses on the detection of frontier model's dangerous capabilities that are not verbalized in their Chain-of-Thoughts (CoTs). Externalized reasoning oversight is a safety paradigm focusing on monitoring the CoTs of large language models (LLMs). As frontier models learn that they are monitored, they can become incentivized to hide their dangerous capabilities from their CoTs. Moreover, as these models become increasingly capable, their obfuscation capabilities also increase, allowing them to bypass monitors. This project investigates ways to detect and defend against such emerging scenarios.

We strive in publishing in top-tier conferences and journals. Specifically, you will work on:

# IMPERIAL

- Designing and developing novel algorithms for detecting and defending against un verbalized dangerous capabilities in frontier models.
- Conducting independent and collaborative research within the group.
- Publishing results in top-tier conferences and journals (e.g., NeurIPS, ICML, ICLR, ACL, EMNLP, AIJ, TACL, etc).
- Implementing research ideas.
- Contributing to research projects involving frontier models.
- Contributing to the life and development of the Safety & Alignment lab, such weekly meetings, presentations, etc.
- Mentoring PhD and MSc students.
- Presenting research at international conferences and workshops.
- Contributing to grant proposals and collaborative research initiatives.

## Key Responsibilities

Main Duties:

Research Duties:

- To take initiatives in the planning and development of research
- To direct the work of small research teams
- To write reports for submission to research sponsors
- To present findings to colleagues and at conferences
- To submit publications to refereed conferences and journals
- To provide guidance to staff and students
- To attend relevant workshops and conferences as necessary
- To develop contacts and research collaborations within the College and the wider community
- To promote the reputation of the group, the Department and the College
- Contribute to bids for research grants
- To conduct and plan own scientific work with appropriate supervision.
- To maintain highly organised and accurate record of experimental work
- To participate in Group research meetings and internal seminars
- To collaborate with colleagues within Imperial College and elsewhere in London and abroad, as appropriate
- Assist in the supervision of UG and PG students and research assistants as required
- To comply with the College, Division, and Unit safety practices and to attend courses on safety when appropriate
- Any other duties as may be deemed reasonable by Head of group as well as Head of Division/Department/Section

Other Duties:

- To undertake any necessary training and/or development.
- To attend relevant meetings.

Where Imperial or funder conditions necessitate, you will be required to complete timesheets for your work on projects in a timely manner.

**About Imperial.** Imperial College is globally recognized, consistently ranking among the top 10 universities worldwide and holding high accolades for research excellence. The Department of Computing is rated 1st in Research Intensity and has been recognized as "Excellent" in the national assessment of teaching quality. This is

# IMPERIAL

your opportunity to be part of a world-leading institution driving forward computational sciences.

## Person Specification

| Requirements                                                           | Essential(E)/<br>Desirable (D) |
|------------------------------------------------------------------------|--------------------------------|
| Candidates/post holders will be expected to demonstrate the following: |                                |

### Education

|                                                                                                                                                                                                                                                       |   |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| Hold a PhD in Computer Science, Mathematics, Physics or a closely related discipline, or equivalent research/industrial experience.<br><i>*Candidates who have not yet been officially awarded their PhD will be appointed as Research Assistant.</i> | E |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|

### Experience

|                                                                                                                                                                             |   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| Practical experience within a research environment and publications in relevant and refereed journals/conferences, such as NeurIPS, ICML, ICLR, ACL, EMNLP, TACL, AIJ, etc. | E |
| Experience of dealing with industry partners and research collaborators                                                                                                     | D |
| Outstanding research experience in frontier models                                                                                                                          | E |
| Practical experience in a broad range of techniques including model checking, monitoring, and CoT                                                                           | E |
| Experience in coding with deep learning libraries such as Pytorch                                                                                                           | E |

### Knowledge

|                                                                                   |   |
|-----------------------------------------------------------------------------------|---|
| Knowledge of programming proficiency in frameworks, such as PyTorch or TensorFlow | E |
| Knowledge of research methods, experimentation and statistical procedures         | E |

### Skills & Abilities

|                                                           |   |
|-----------------------------------------------------------|---|
| Ability to conduct a detailed review of recent literature | E |
|-----------------------------------------------------------|---|

# IMPERIAL

|                                                                                                            |   |
|------------------------------------------------------------------------------------------------------------|---|
| Ability to develop and apply new concepts                                                                  | E |
| Creative approach to problem-solving                                                                       | E |
| Excellent verbal communication skills and the ability to deal with a wide range of people                  | E |
| Excellent written communication skills and the ability to write clearly and succinctly for publication     | E |
| Ability to organise own work with minimal supervision                                                      | E |
| Ability to prioritise own work in response to deadlines                                                    | E |
| Ability to direct the work of a small research team and motivate others to produce a high standard of work | E |
| Willingness to work as part of a team and to be open-minded and cooperative                                | E |
| Flexible attitude towards work                                                                             | E |
| Discipline and regard for confidentiality and security at all times                                        | E |
| Willingness to undertake any necessary training for the role                                               | E |
| Willingness to travel both within the United Kingdom and abroad to attend conferences                      | E |

## Further Information

Please note that job descriptions cannot be exhaustive, and the post-holder may be required to undertake other duties, which are broadly in line with the above key responsibilities.

[Our values](#) are at the root of everything we do and everyone in our community is expected to demonstrate Imperial:

- Respect
- Collaboration
- Excellence
- Integrity
- Innovation

Employees are also required to comply with all Imperial policies and regulations.

We are committed to equality of opportunity, to eliminating discrimination and to creating an inclusive working environment for all. We encourage candidates to apply irrespective of age, disability, marriage or civil partnership status, pregnancy or maternity, race, religion and belief, gender reassignment, sex, or sexual orientation. You can read more about our commitment [on our webpages](#).

Imperial is a proud signatory to the San-Francisco Declaration on Research Assessment (DORA), which means that in hiring and promotion decisions, we evaluate applicants on the quality of their work, not the journal impact factor where it is published. For more information, see <https://www.imperial.ac.uk/research-and-innovation/about-imperial-research/research-evaluation/>

Imperial believes that the use of animals in research is vital to improve human and animal health and welfare. Animals may only be used in research programmes which are ultimately aimed towards finding new treatments and making scientific and medical advances, and where there are no satisfactory or reasonably practical alternatives to their use. Imperial is committed to ensuring that, in cases where this

# IMPERIAL

research is deemed essential, all animals in our care are treated with full respect, and that all staff involved with this work show due consideration at every level. [Find out more about animal research at Imperial.](#)

October 2026