



Job Description

Summary

Job title	Research Assistant on AI Safety
Division	Mathematical, Physical, and Life Sciences (MPLS)
Department	Computer Science
Location	Wolfson Building, Parks Road, Oxford
Grade and salary	£35,681- £39,424 per annum. This is inclusive of a pensionable Oxford University Weighting of £1,730 per year.
Hours	Full time
Contract type	Fixed-term for 6 months with possible extension
Reporting to	Prof Yarin Gal
Vacancy reference	188116
Additional information	1 post

Research topic	Generative AI, AI Safety
Principal Investigator	Prof Yarin Gal
Funding partner	The funds supporting this research project are provided by Coefficient Giving (formerly Open Philanthropy)

The role

The Department of Computer Science is looking to employ a Research Assistant to work within a research group, reporting to Prof Yarin Gal and Dr Lin Li. The postholder will be a member of the Oxford Applied and Theoretical Machine Learning (OATML) research group with responsibility for the provision of research support for projects relating to AI technical safety, security, reliability and robustness, as well as foundational research on frontier models like Large Language Models (LLMs) and multimodal foundation models. This includes topics such as safety alignment, adversarial training, jailbreaking, uncertainty quantification, and AI hallucination.

As a research assistant, you will contribute to projects aimed at developing principled and practical methods which could be used in real systems. The project will involve both theoretical work as well as empirical analysis on challenging tasks.



Responsibilities

- Carry out research activities within guidelines provided by Professor Yarin Gal and Dr Lin Li.
- Contribute to wider project planning, including ideas for new research projects.
- Select, follow, and adapt specialist methodologies and identify suitable alternatives where information or research material is restricted.
- Contribute to the design of research materials.
- Gather, analyse, and present qualitative and/or quantitative data from a variety of sources.
- Contribute to research publications, book chapters and reviews.
- Support collaboration activities with external stakeholders, as directed by Professor Yarin Gal and Dr Lin Li.
- Contribute to discussions and share research findings with colleagues in partner institutions, and research groups.

Selection criteria

- Hold a relevant Master's degree (or be close to completion¹).
- Demonstrated expertise and familiarity with literature and research in Foundation Models (generative AI), LLMs jailbreaking, safety alignment, adversarial machine learning or general deep learning.
- Strong programming skills.
- Possess sufficient specialist knowledge in the discipline to contribute to development of research methodologies as well as develop, plan and manage research projects.
- Excellent communication skills, including the ability to write for publication, present research proposals and results and represent the research group at meetings and conferences.

¹ Evidence required (please provide one of the following):

- A copy of your Master's degree certificate; or
- An academic reference confirming that your Master's qualification has been awarded; or
- If your degree has not yet been awarded, an academic reference confirming that you have submitted your dissertation.

Desirable

- Experience with large language models related to jailbreaking, safety alignment, and adversarial machine learning.
- Publication record in Foundation Models (generative AI), LLMs jailbreaking, safety alignment, adversarial machine learning or general deep learning.

Pre-employment screening

Standard checks

If you are offered the post, the offer will be subject to standard pre-employment checks. You will be asked to provide: proof of your right-to-work in the UK; proof of your identity; and (if we haven't done so already) we will contact the referees you have nominated. If you have previously worked for the University we will also verify key information such as your dates of employment and reason for leaving your previous role with the department/unit where you worked. You will also be asked to complete a health declaration so that you can tell us about any health conditions or disabilities for which you may need us to make appropriate adjustments.

Please read the candidate notes on the University's pre-employment screening procedures at:

<https://www.jobs.ox.ac.uk/pre-employment-checks>

About the University of Oxford

Welcome to the University of Oxford. We aim to lead the world in research and education for the benefit of society both in the UK and globally. Oxford's researchers engage with academic, commercial and cultural partners across the world to stimulate high-quality research and enable innovation through a broad range of social, policy and economic impacts.

We believe our strengths lie both in empowering individuals and teams to address fundamental questions of global significance, while providing all our staff with a welcoming and inclusive workplace that enables everyone to develop and do their best work. Recognising that diversity is our strength, vital for innovation and creativity, we aspire to build a truly diverse community which values and respects every individual's unique contribution.

While we have long traditions of scholarship, we are also forward-looking, creative and cutting-edge. Oxford is one of Europe's most entrepreneurial universities and we rank first in the UK for university spin-outs, and in recent years we have spun out 15-20 new companies every year. We are also recognised as leaders in support for social enterprise.

Join us and you will find a unique, democratic and international community, a great range of staff benefits and access to a vibrant array of cultural activities in the beautiful city of Oxford.

For more information, please visit www.ox.ac.uk/about/organisation.

Department of Computer Science

The Department of Computer Science is consistently recognised as the internationally leading centre of research and teaching across a broad spectrum of computer science. We are proud of our history as one of the longest-established computer science departments in the country, as we continue to provide first-rate undergraduate and postgraduate teaching to some of the world's brightest minds.

Our world-class research is conducted across our research themes, which span the broad spectrum of computer science, ranging from foundational discoveries to interdisciplinary work with significant real-world impact. A significant majority of our staff are active in externally sponsored research, with both government and industrial funding. Our 2021 Research Excellence Framework submission saw 81% of our research activity ranked as world-leading (4*), with the rest ranked as internationally excellent (3*). We have had 19 ERC Fellowships in the last

decade (including 7 Advanced) and we have 6 Fellows of the Royal Society, 4 Turing/Turing AI Fellows, a Fellow of the Royal Academy of Engineering and a Fellow of the Institute of Electrical and Electronics Engineers.

We enjoy close links with other Oxford University departments (Mathematics, Engineering, Physics, Statistics and Life sciences) and work collaboratively with Oxford research groups and institutes (including the Oxford Internet Institute and the Oxford e-Research Centre). At present, the department has 71 faculty members and 98 researchers housed across multiple sites within the University's South Parks Road Science Area and the neighbouring area. Through a programme of continuous improvement, the department is committed to promoting and nurturing a diverse, inclusive and equal culture, with a particular focus on growth in gender equality (from our students to our staff).

The department holds over £75m of external funding of which £58m is research. Research in the department is currently managed in ten themes:

- *Algorithms & Complexity Theory*, led by Professor Leslie Ann Goldberg, focusses on determining the inherent difficulty of computational problems, classifying problems according to this inherent difficulty, and designing and analysing algorithms that use computational resources as efficiently as possible;
- *Artificial Intelligence & Machine Learning*, led by Professor Michael Wooldridge, focuses on theoretical foundations of AI, multi-agent systems, deep learning, reinforcement learning, and computational linguistics;
- *Automated Verification*, led by Professor Marta Kwiatkowska, investigates theory and practice of formal verification and correct-by-construction synthesis for software and hardware systems;
- *Computational Biology & Health Informatics*, led by Professor Blanca Rodriguez, is concerned with computational approaches for biomedical research and healthcare innovation;
- *Human-Centred Computing*, led by Professor Nigel Shadbolt, includes human-computer interaction, social computing, and the worldwide web;
- *Data and Knowledge & Action*, led by Professor Ian Horrocks, includes databases, knowledge representation and reasoning;
- *Programming Languages*, led by Professor Nobuko Yoshida, includes functional programming, program analysis, and programming language foundations;
- *Quantum*, led by Professor Jonathan Barrett, focusses on quantum computing including quantum software, causality in quantum theory, quantum cryptography and foundations of quantum computing;
- *Security*, led by Professor Ivan Martinovic, specialises in cybersecurity, protocol analysis, systems security, trusted computing, and networking.
- *Systems*, led by Professor Niki Trigoni, focusses especially on cyber physical systems. We plan to substantially broaden our research in systems to complement our existing research areas.

Our greatest asset is our people. We consistently attract the best staff and students, and thanks to them, we have been ranked as the world's leading university for Computer Sciences for eight years in a row by the *Times Higher Education*. We have held an Athena Swan Bronze Award since 2014, reflecting our longstanding commitment to promoting and supporting gender equality.

Find out more information on our website <https://www.cs.ox.ac.uk/>

Mathematical, Physical, and Life Sciences Division

The Mathematical, Physical, and Life Sciences (MPLS) Division is one of the four academic divisions of the University. Oxford is widely recognised as one of the world's leading science universities and the MPLS Division is home to our non-medical sciences, with 9 academic departments that span the full spectrum of the mathematical, computational, physical, engineering and life sciences, and undertake both fundamental research and cutting-edge applied work.

For more information about the MPLS division, please visit: www.mpls.ox.ac.uk

How to apply

Applications are made through our online recruitment portal. Information about how to apply is available on our Jobs website <https://www.jobs.ox.ac.uk/how-to-apply>.

Your application will be judged solely on the basis of how you demonstrate that you meet the selection criteria stated in the job description.

As part of your application you will be asked to provide details of two referees and indicate whether we can contact them now.

You will be asked to upload a CV and a supporting statement. The supporting statement must explain how you meet each of the selection criteria for the post using examples of your skills and experience. This may include experience gained in employment, education, or during career breaks (such as time out to care for dependants)

Please upload all documents **as PDF files** with your name and the document type in the filename.

All applications must be received by **midday** UK time on the closing date stated in the online advertisement.

Information for internal candidates

If you currently work at the University please note that:

- as part of the referencing process, we will contact your current department to confirm basic employment details including reason for leaving and information about your performance.
- although employees may hold multiple part-time posts, they may not hold more than the equivalent of a full time post. If you are offered this post, and accepting it would take you over the equivalent of full-time hours, you will be expected to resign from, or reduce hours in, your other posts(s) before starting work in the new post.
-

Information for priority candidates

A priority candidate is a University employee who is seeking redeployment because they have been advised that they are at risk of redundancy, or on grounds of ill-health/disability. Priority candidates are issued with a redeployment letter by their employing department(s).

If you are a priority candidate, please ensure that you attach your redeployment letter to your application (or email it to the contact address on the advert if the application form used for the vacancy does not allow attachments).

If you need help

Application FAQs, including technical troubleshooting advice is available at:

<https://staff.web.ox.ac.uk/recruitment-support-faqs>

Non-technical questions about this job should be addressed to the recruiting department directly hr@cs.ox.ac.uk

To return to the online application at any stage, please go to: www.recruit.ox.ac.uk.

Please note that you will receive an automated email from our online recruitment portal to confirm receipt of your application. **Please check your spam/junk mail** if you do not receive this email.

Important information for candidates

Data Privacy

Please note that any personal data submitted to the University as part of the job application process will be processed in accordance with the GDPR and related UK data protection legislation. For further information, please see the University's Privacy Notice for Job Applicants at: <https://compliance.admin.ox.ac.uk/job-applicant-privacy-policy>. The University's Policy on Data Protection is available at: <https://compliance.admin.ox.ac.uk/data-protection-policy>.

The University's policy on retirement

The University operates an Employer Justified Retirement Age (EJRA) for very senior research posts at **grade RSIV/D35 and clinical equivalents E62 and E82** of 30 September before the 70th birthday. The justification for this is explained at: <https://hr.admin.ox.ac.uk/the-ejra>.

For **existing** employees on these grades, any employment beyond the retirement age is subject to approval through the procedures: <https://hr.admin.ox.ac.uk/the-ejra>.

There is no normal or fixed age at which staff in posts at other grades have to retire. Staff at these grades may elect to retire in accordance with the rules of the applicable pension scheme, as may be amended from time to time.

Equal Opportunity

The University of Oxford is committed to equal opportunity, and to being a place where everyone belongs and is supported to succeed. We recognise how the diversity of our community enriches our ability to deliver on our academic mission.

We welcome applications from individuals from all backgrounds, including those under-represented within higher education. No applicant or members of staff shall be unlawfully discriminated against on the basis of age, disability, gender reassignment, marriage or civil partnership, pregnancy or maternity, race, religion or belief, sex, or sexual orientation.

Employment with the University and progression within employment will be determined according to personal merit and the application of criteria related to the duties and conditions of the post. In all cases, the primary consideration will be the ability to perform the job.

As stated in the University's Equality Policy and Equality, Diversity and Inclusion Strategic Plan, our commitment to equality and diversity goes hand in hand with our commitment to academic freedom and free speech.

Benefits of working at the University

Employee benefits

University employees enjoy 38 days' paid holiday, generous pension schemes, flexible working options, support for [sustainable travel](#) and other discounts. Staff can also access a huge range of personal and professional development opportunities. See <https://hr.admin.ox.ac.uk/staff-benefits>. Staff can access a huge range of personal and professional development opportunities. See <https://hr.admin.ox.ac.uk/staff-benefits>

Employee Assistance Programme

As part of our wellbeing offering staff get free access to a confidential employee assistance programme, available 24/7 for 365 days a year. Find out more at <https://staff.admin.ox.ac.uk/thriving-at-oxford>

University Club and sports facilities

Membership of the University Club is free for University staff. It offers social, sporting, and hospitality facilities. Staff can also use the University Sports Centre on Iffley Road at discounted rates, including a fitness centre, powerlifting room, and swimming pool. See www.club.ox.ac.uk and <https://www.sport.ox.ac.uk/>.

Information for staff new to Oxford

Please see our Life in Oxford webpage for information on relocating to and settling into the Oxford area. The website offers valuable guidance, including information on where to find more details about housing, transportation, finances, healthcare, and other key aspects of living in Oxford and the surrounding region. See [Life in Oxford | Oxford University Jobs](#)

There is also a visa loan scheme to cover the costs of UK visa applications for staff and their dependants. See <https://staffimmigration.admin.ox.ac.uk/visa-loan-scheme>

Family-friendly benefits

We are a family-friendly employer with one of the most generous family leave schemes in the Higher Education sector (see <https://hr.web.ox.ac.uk/family-leave>). Our Childcare Services team provides guidance and support on childcare provision, and offers a range of high-quality childcare options at affordable prices for staff. In addition to 5 University nurseries, we partner with a number of local providers to offer in excess of 450 full time nursery places to our staff. Eligible parents are able to pay for childcare through salary sacrifice, further reducing costs. See <https://childcare.admin.ox.ac.uk/>.

Supporting disability and health-related issues (inc menopause)

We are committed to supporting members of staff with disabilities or long-term health conditions, including those experiencing negative effects of menopause. Information about the University's Staff Disability Advisor, is at <https://edu.admin.ox.ac.uk/disability-support>. For information about how we support those going through menopause see <https://hr.admin.ox.ac.uk/menopause-guidance>

Staff networks

The University has a number of staff networks including for research staff, BME staff, LGBT+ staff, disabled staff network and those going through menopause. Find out more at <https://edu.admin.ox.ac.uk/networks>

The University of Oxford Newcomers' Club

The University of Oxford Newcomers' Club is run by volunteers that aims to assist the partners of new staff settle into Oxford, and provides them with an opportunity to meet people and make connections in the local area. See www.newcomers.ox.ac.uk.

Research staff

The Researcher Hub supports all researchers on fixed-term contracts. They aim to help you settle in comfortably, make connections, grow as a person, extend your research expertise and approach your next career step with confidence. Find out more <https://www.ox.ac.uk/research/support-researchers/researcher-hub>

Oxford's Research Staff Society is a collective voice for our researchers. They also organise social and professional networking activities for researchers. Find out more <https://www.ox.ac.uk/research/support-researchers/connecting-other-researchers/oxford-research-staff-society>