

DR SARVNAZ KARIMI

Principal Research Scientist
CSIRO Data61
Marsfield NSW 2122, Australia
Phone: +61 2 9372 4353
Email: sarvnaz.karimi@csiro.au

Bio

Sarvnaz is a principal research scientist at CSIRO's Data61, Language Technology Team. She has nearly two decades of experience working in NLP (Natural Language Processing) and Search and Information Retrieval (IR) domains. She applies her expertise in Information Retrieval, Natural Language Processing, and Machine Learning intersection on different real-world projects in domains such as health, climate adaptation and agriculture. She has led multiple projects in the NLP/IR space over the years.

Sarvnaz is active in the research community in Australia and internationally. She is currently the Director of Publicity at the Association for Computational Linguistics (ACL), NLP's main international professional body. She is one of the editors-in-chief of the ACL Rolling Review, the centralised reviewing service targeting top-tier conferences under the Association for Computational Linguistics. She's on the advisory board of Widening NLP, an ACL-affiliated affinity group dedicated to increasing the visibility and participation of researchers from underrepresented groups within the NLP, CL, and AI communities (2025–2027). Sarvnaz acted as the President of the Australasian Language Technology Association (ALTA), and an executive committee member (8 years in total). She has served as chair, senior area chair, area chair, and Programme Committee member of several major national and prestigious international conferences and journals, including the Association for Computing Machinery's Special Interest Group on Information Retrieval (SIGIR), the Conference on Information and Knowledge Management (CIKM), the International Conference on Computational Linguistics (COLING), and the Association for Computational Linguistics (ACL), and Transactions of the Association for Computational Linguistics (TACL). Through CSIRO and the Data61 scholarship programme, Sarvnaz co-supervises PhD students from different Australian universities.

Sarvnaz completed a PhD in Natural Language Processing in 2008 (RMIT University). She has a masters in Artificial Intelligence and a bachelors degree in Computer Science and Software Engineering.

Homepage: people.csiro.au/K/S/Sarvnaz-Karimi
LinkedIn: www.linkedin.com/in/sarvnaz-karimi-64ab491
Google Scholar (Publications): [goo.gl/KoTmwz](https://scholar.google.com/citations?user=KoTmwz)

Research Interests

Natural Language Processing, Computational Linguistics, Search and Information Retrieval, Digital Health, Biomedical Text Mining, Social Media Analytics, NLP for Cli-

mate Science and Agriculture.

Education

Feb 2005 – Mar 2008 PhD in Computer Science (Natural Language Processing)
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia
Research topic: Machine transliteration/Translation

Professional Experience

Mar 2012 – present *Principal Research Scientist*
CSIRO Data61, Marsfield, NSW

- Research member of the Language Technology Team, Analytics and Decision Sciences Programme.
- Jun—July 2024: Acting Group Leader of Humans and Machines Group (50 people)
- Oct 2017—Nov 2017: Acting Team Leader of Language Technology Team (during this time the team grew by 5 new postdoctoral fellows, and 5 interns, taking care of hiring processes including forming the interview team, the content of the interview, and leading the panels)
- Aug 2023—Apr 2024: Acting Team Leader of Language Technology Team
- Dec 2019—Jun 2020: Chair of NSW/SA CSIRO Data61 PhD scholarship committee. It involved assessing all the scholarship applicants, chairing the assessment sessions and finalising a recommendation of the ranking of the applicants to the CSIRO Data61's Chief Scientist.
- Nov 2017—Nov 2019: Member of CSIRO Data61 PhD scholarship committee for VIC/TAS/WA states. This role involves reviewing PhD scholarship applications and ranking them in a committee meeting three to four times a year.

Projects:

- Boeing Advanced Production Systems (APS) (2024–present)
The project goal is to design AI/ML tools at scale to advance Boeing advance production systems. Our sub-team investigates the role of NLP technologies in addressing some of the needs in processing their documentation.
- Climate Services for Agriculture (CSA) (2023–present)
As part of the CSA team, doing a feasibility study to identify ways a system using Large Language Models (LLMs) could assist farm advisors in answering questions using reliable information sources, such as scientific literature.
- Science Digital (2023-2025)
NLP Lead for a usecase on genomics, where tools and evaluation frameworks were developed for scientists, with NLP techniques on information extraction and summarisation of biomedical literature, and question answering based on the published articles in the biomedical field.

- NLP for Mental Health (2023–present)
This project is a collaboration between CSIRO and Macquarie University (Prof Maria Kangas, Head of the School of Psychological Sciences and Dr Diego Molla-Aloid, School of Computing). It explores opportunities to develop technology that assists in identifying coping mechanisms for people who show signs of resilience in times of hardship. The project has acquired data from an industry collaborator, Vent, that runs a platform of a similar name, allowing its users to share their daily lives through posts.
- Medical Decision Support (2018–2024)
Leading a Precision Health Future Science Platform project (\$2.2m budget) on improving decision-making for precision medicine. As part of this project, three search prototypes are developed: A2A (search benchmarking), Science2Cure (clinical trials and scientific literature search) and SearcEHER (clinical notes on family histories information extraction and search).
Our ranking algorithms scored among the top 10 teams internationally in the TREC evaluation tasks on precision medicine in 2019 and ranked 1st in the task in 2020 and 2023. Our search algorithm is tailored for COVID search (pandemic retrieval) to improve accessing information in the published literature on coronavirus ranked 4th internationally among 143 submissions. The corresponding publication for this algorithm won the best short paper award in ALTA 2020. Hired and mentored two postdocs as part of this project. Two PhD students, 7 summer interns, and one honours student were trained and graduated as part of this project. The project led to 10 journal publications and 17 peer-reviewed conference papers.
- Syndromic Surveillance for Infectious Diseases (2017–2022)
The goal of this project is to use social media, such as Twitter, to assist in identifying infectious diseases in specific geographical locations. This project is in collaboration with the UNSW Public Health research group at the Kirby Institute. We were successful in attracting an Office of the Chief Executive (OCE) grant to hire a postdoctoral fellow as well as funding from CSIRO Health & Biosecurity's Risk Evaluation and Preparedness Programme. The team has also secured funding from MRFF grants for 2021-2022 (led by UNSW) to further the research. We have published 4 journal papers and 4 conference papers in top venues. Our team was ranked first in an international competition on vaccination behaviour detection from social media. A tutorial for Syndromic Surveillance was delivered at one of the leading computational linguistics conferences in December 2020.
- Low Carb Meals Commercialisation (2019–2023)
Responsible for evaluating the effectiveness of CSIRO's low-carb diet book series and the endorsed meals through social media feedback.
- Deep learning for Clinical Support (2018–2019)
In collaboration with researchers in the CSIRO's eHealth programme, developed algorithms that allowed processing of the clinical documents, such as radiology reports, across multiple hospitals. We collaborated with clinicians from Gold Coast hospitals to evaluate such transferrable systems.
- Search for Clinical Decision Making (2016–2018)
Led this project with two aims: (1) developing a public platform for researchers to

experiment on information retrieval algorithms that advance search for biomedical articles; and, (2) developing algorithms that match cancer patients with clinical trials. This project was the basis for the Precision Health Future Science Platform Project.

- Artificial Intelligence Roadmap (2018–2019)
Part of the team that wrote Artificial Intelligence Roadmap: Solving problems, growing the economy and improving our quality of life. This report was requested and delivered to the Australian Government.
- Information extraction for Digital Leasing (Lease Info) (2017)
Providing consultancy and a proof-of-concept product, in the domain of information extraction, for an SME in Sydney.
- Australia At Your Service (2016)
An expert search system, Network (network.csiro.au), was developed and is currently in use internally within CSIRO. This project also led to some involvement in another expertise search project called ExpertConnect (<https://expertconnect.data61.csiro.au/>).
- Adverse Drug Reaction Discovery from Patient Reports (2013–2015)
Led a two-year project to (1) mine patient forums for reports of adverse drug events; and (2) find duplicate reports of adverse events in the Therapeutic Goods Administration (TGA) spontaneous reporting system. This project led to highly cited publications in journals such as ACM Computing Surveys and the Journal of Biomedical Informatics, as well as two software prototypes. We also released a dataset called CADEC through a highly cited journal paper (over 300 citations) that is internationally used for developing similar methods in the biomedical text-mining community.
- Enhanced coding of cause of death recorded on death certificates (2015–2016)
Applied text classification methods on free-text reports of death certificates to automatically assign ICD10 codes (International Coding of Diseases). The outcome was a system that is currently in use in the NSW Health influenza monitoring system.
- Emergency Situation Awareness (2012–2013)
A research member of the team that provided NLP expertise and developed methods to identify social media posts indicating different categories of disasters and developed geolocation prediction methods for Twitter messages. During this time, I also co-supervised two interns and one honours thesis from the University of Sydney.

May 2008 – Mar 2012 *Researcher*

NICTA Victoria Research Laboratory, The University of Melbourne, VIC
Research member of the Biological Text and Language Applications (BioTALA) project, Making Sense of Data (MSD) theme.

In BioTALA, I developed techniques for processing biomedical text. Some of these activities were focused on the specific needs of external collaborators (e.g., The Global Evidence Mapping Initiative and Alfred Hospital).

My main research centred around developing algorithms for improving search in biomedical text, including search for relevant published research for the creation of clinical systematic reviews and genomics information retrieval. I worked on extracting drug side effects from public forums and searching in medical records.

Semester 2, 2011 Guest Lecturer (Information retrieval)
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia

Feb 2005 – Oct 2007 *Tutor*
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia

- Database Concepts (5 semesters)
- Web Database Applications (2 semesters)
- Scripting Language Programming (Perl and Python)
- File Structures and Database Systems (2 semesters), Head Tutor
- Web Programming
- Web Servers and Web Technology
- Computer and Internet Forensics
- Secure Electronic Commerce
- Unix Essentials and C shots (RMIT Orientation Programme, four semesters)

September 2005 – November 2005 *Online Instructor*
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia
Introduction to Information Technology

Aug 1999 – Jan 2005 *Research Engineer*
Iran Telecommunication Research Centre (ITRC), Tehran, Iran

- Member of text mining research group (October 2002 – January 2005)
Responsibilities: Improving the performance of TelQAS, an ontology-based telecommunication question answering system developed in-house, using query expansion (pseudo-relevance feedback), document clustering, and multi-document summarisation.
- Member of network management research group (August 1999 – September 2002)
Responsibilities: Improving basic call service model in Intelligent Networks (IN)
Real-time services in Internet, Voice Over IP (VOIP)
TMN (Telecommunication Management Network) in mobile networks

Honours, Awards, and Grants

The Global Science and Technology Diplomacy Fund Strategic Element (GSTDF-SE), shortlisted in August 2025 for a \$1m project on Climate Adaptation with three part-

ner countries of Malaysia, Indonesia and Vietnam. Outcome of the final round to be announced in October 2025.

CSIRO Excellence in Project Delivery Award, 2025 (for my role as science lead of Boeing APS Project)

Best runner-up paper award at CIKM 2023 Main Conference (Core A Ranking)

Medical Research Future Fund (MRFF) 2021 Frontier Health and Medical Research Initiative (Led by UNSW Kirby Institute, EpiWatch Artificial Intelligence Early-Warning System for Epidemics), with the value of \$799,788.

Best Short Paper Award, The 2020 Australasian Language Technology Association (ALTA) 2020.

CSIRO non-cash award in recognition of praiseworthy achievement in the Precision Medicine 2020 International Challenge (top ranking system).

Outstanding reviewer award from the Empirical Methods in Natural Language Processing (EMNLP) Conference in 2018 and 2019 (CORE ranking A conference in Natural Language Processing).

Part of the *LIS-Accurait* project team that won two merit awards at the 2019 NSW iAwards ceremony for *Infrastructure and Platforms Innovation of the Year* and *Business Service Markets*.

OCE Postdoctoral Fellowship grant on “Improving the detection, diagnosis and prognosis of infectious diseases by integrating current syndromic. surveillance tools with social media insights”, with a value of \$105K per year for three years (2018–2021), together with Dr Ross Spark, Dr Cecile Paris, Dr Paul De Barro, and Prof. Raina McIntyre (UNSW). This allowed us to hire a postdoc under our supervision.

CSIRO non-cash award in recognition of praiseworthy achievement in supporting “The Power of Peoples Data” stand at D61+Live 2019.

CSIRO Program Recognition Rewards in Excellence, 2017. As part of the COLING team that won an international competition to identify the location of Twitter posts.

CSIRO Program Recognition Rewards in Performance, 2017. As part of the NETWORK team which negotiated CSIRO’s social, institutional and technical complexity to develop and launch a product to improve CSIRO’s capability discoverability in a very short space of time (<http://network.csiro.au>).

Commendation on the PhD thesis to be recognised as a work that is *of the highest merit and at the forefront internationally in its (examiners: A/Prof James Thom, Prof. Kevin Knight, A/Prof Ari Pirkola). field*, by the international thesis examiners, 2008.

Best full paper award, The 14th Australasian Document Computing Symposium (ADCS’09), Sydney, Australia, December 2009.

Member of the team that became a semi-finalist at the Elsevier Grand Challenge (from 70 entries): Timothy Baldwin, Lawrence Cavedon, Sarvnaz Karimi, David Martinez,

David Newman, Falk Scholer and Justin Zobel, “Effective Search, Classification, and Visualisation of Information from Large Collections of Biomedical Literature”.

Best presentation award, RMIT Research Student Conference (CRSC), 2006.

Student travel grant (£500) to attend String Processing and Information Retrieval (SPIRE), 2006, Glasgow, UK.

Ranked 15th, the Iran trans-country university entrance exam (for masters degree), 2001.

Top student, gold coin recipient (2nd place among the university graduates of BEng, Computer Engineering), 2000.

University student research prize, 1999.

Member of the university ACM programming team which ranked 7th in Tehran site, West Asia region of the ACM international collegiate programming contest, 1999.

Talks, Media & Panels

Invited talk, UNSW BioMedical Machine Learning Lab, Title: “Measuring summarisation quality of Large Language Models on scientific articles” (March 2025)

Invited talk, Google Developers Group, Co-presented with the CSIRO colleagues and Google partners on AI for Science program (November 2024)

Panelist, How Are Emerging Research Trends Shaping a Changing Academic Landscape?, AI and Cyber Futures Institute annual research conference (November 2024)

Panelist, AI for Science (A Google and CSIRO co-hosted community discussion), 2024 eResearch Australasia Conference (October 2024)

Commoditising the use of AI for scientists: How AI-agents assist and transform scientific discovery, co-presented with Dr Q. Lu and Dr S. Harrar, 2024 eResearch Australasia Conference (October 2024)

Presentation Matters: How to Communicate Science in the NLP Venues and in the Wild?, 3-hour tutorial co-presented with Dr C. Paris and Prof G. Haffari, ACL 2024 Conference (August 2024)

Panelist at “Building a strong AI presence for SMEs to make the best of what technology can offer securely”, AI and Cyber for SME Growth Symposium 2021, 22 June 2021.

Media report by Medical Republic, February 2021
Intelligent clinical trials.

Media report by Medical Republic, January 2020
Why health is the ‘worst case’ scenario for text mining.

Tutorial Speaker at Australasian Language Technology Association 2019 Tutorial, December 2019
Title: *NLP for Healthcare in the Absence of a Healthcare Dataset*, co-presented with Aditya Joshi.

Speaker at CSIRO Connect 2019, August 2019

Title: *AI for Health*.

Speaker for Big Day of Data61 Scholars, February 2019

Title: *Your social media and our research*.

Panelist at Sydney Natural Language Processing Meetup, April 2018

Panel topic: *The next 30 years of information extraction*.

Invited speaker at CMCRC (Capital Markets CRC), July 2017

Title: *Information extraction and retrieval for decision making in health and precision medicine*.

Invited talk at Australian Institute of Health and Welfare, 2017

Title: *Textual Analysis: Techniques, New Data, Opportunities and Challenges*, co-presented with Cecile Paris.

Invited talk at PwC (PriceWaterhouseCoopers) Australia, 2015

Presented on behalf of Scientists & Mathematicians in School to share my experience as a volunteer scientist.

Invited speaker at ARCS Australia Scientific Congress (<https://www.arcs.com.au/>), ARCS Pharmacovigilance and Social Media, June 2014

Title: *Text Mining for Pharmacovigilance*, presented at Pharmacovigilance and Social Media Session.

Invited talk at Amsterdam University, Information Retrieval Group, 2011

Title: *Systematic Reviews for Evidence-Based Medicine*.

Invited talk at the monthly meetings of Linux users of Victoria (LUV), The University of Melbourne, 1st June 2010.

Title: *Biomedical Search Engines*

Certificates

Leadership with Impact, Leading Lighter, 2024

Core Leadership, Leading Lighter, 2024

Leadership Essentials, CSIRO, 2014

Project Management Fundamentals, CSIRO, 2013

Publications

A full list of publications and their citations can be found on my Google Scholar page: [goo.gl/K0Tmwz](https://scholar.google.com/citations?user=K0Tmwz). Some of my journal articles are listed below: (IF: Impact Factor)

- Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis M. Rybinski, X. Dai, S. Singh, S. Karimi, A. Nguyen, JMIR medical informatics, 9 (4), e24020, 2021 (IF: 2.58).
- Show, tell and summarise: learning to generate and summarise radiology findings

from medical images S Singh, S Karimi, K Ho-Shon, L Hamey, Neural Computing and Applications, 1-25, 2021 (IF: 5.40).

- A2A: a platform for research in biomedical literature search, M Rybinski, S Karimi, V Nguyen, C Paris, BMC bioinformatics, 21 (19), 1-27, 2020 (IF: 3.242).
- W. Tay, J. Zhang, S. Karimi, Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness, Journal of the Association for Information Science and Technology, pages 784-799, 71(7), 2020. (CORE ranking A*, IF: 2.835)
- H. Hassanzadeh, S. Karimi, A. Nguyen, Matching Patients to Clinical Trials Using Semantically Enriched Document Representation, Journal of Biomedical Informatics, 2020. (IF: 2.95)
- A. Joshi, R. Sparks, J. McHugh, S. Karimi, C. Paris, CR. MacIntyre, Harnessing tweets for early detection of an acute disease event, Epidemiology, 31 (1), 2020. (IF: 2.26)
- A. Joshi, S. Karimi, R. Sparks, C. Paris, CR. MacIntyre, Survey of text-based epidemic intelligence: A computational linguistic perspective, ACM Computing Surveys, 52(6), Article 119. (CORE ranking A*, IF: 6.131)
- B. Koopman, S. Karimi, A. Nguyen, R. McGuire, D. Muscatello, M. Kemp, D. Truran, M. Zhang, S. Thackway, Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Medical Informatics and Decision Making, 15(53), 2015. (IF: 2.067)
- S. Karimi, J. Yin, J. Baum, Evaluation Methods for Statistically Dependent Text, Computational Linguistics, pages 539-548, 41(3), 2015. (CORE ranking A*, IF: 1.319)
- S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, CADEC: A Corpus of Adverse Drug Event Annotations, Journal of Biomedical Informatics. pages 7381, Vol. 55, June 2015. (IF: 2.95)
- S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Text and Data Mining Techniques in Adverse Drug Reaction Detection, ACM Computing Surveys, Vol. 47, Number 4, Article 56, 2015 (CORE ranking A*, IF: 6.131)
- S. Karimi, J. Zobel, F. Scholer, Quantifying the Impact of Concept Recognition on Biomedical Information Retrieval, Information Processing and Management, pages 94-106, 48(1), 2012. (CORE ranking A, 3.892)
- S. Karimi, F. Scholer, A. Turpin, Machine Transliteration Survey, ACM Computing Surveys, 57 pages, 43(3), 2011. (CORE ranking A*, IF: 6.131)
- S. Karimi, S. Pohl, F. Scholer, L. Cavedon, J. Zobel, Boolean versus Ranked Querying for Biomedical Systematic Reviews, BMC Medical Informatics and Decision Making, 10(58), 2010. (IF: 2.067)

Science Citizenship

Note: Computer science conferences generally provide peer reviews for the submitted papers. Each *area chair* is in charge of a subset of submitted manuscripts for which they assign reviewers, moderate a discussion phase, and provide tentative decisions which are handed to technical program chairs.

Chair/Co-chair/Area chair/Editor:

- ACL Rolling Review, Editor-in-Chief (EiC), (2024–present) – note EiCs are allocated as guest program chairs of the ACL conferences they support
- Guest Program Chair of EMNLP 2025 (<https://2025.emnlp.org/organization/>)
- The best paper committee member for EMNLP 2025 (core A* conference)
- Widening NLP Advisory Board Member (2025–present)
- Guest editor-in-chief of an upcoming special issue in the ACM Transactions on Interactive Intelligent Systems (2024-2025), with A/Prof Y. Deldjoo, Dr B. Kocaballi, and Prof. Shlomo Berkovsky
- Senior Area chair of ACL 2023 (The 61st annual meeting of the Association for Computational Linguistics)—Resources and Evaluation track
- ACL Rolling Review (ARR), Action Editor/Area Chair (2020-2024)
- Area chair of EMNLP 2021 (The 2021 Conference on Empirical Methods in Natural Language Processing)—Information Retrieval and Text Mining track
- Area chair of ACL-IJCNLP 2021 (The 58th annual meeting of the Association for Computational Linguistics)—Machine Learning for NLP track
- Area chair of EMNLP 2020 (The 2020 Conference on Empirical Methods in Natural Language Processing)—Machine Learning for NLP track
- Area chair of ACL 2020 (The 58th annual meeting of the Association for Computational Linguistics)—Resources and Evaluation track
- Area chair of ACL 2019 (The 57th annual meeting of the Association for Computational Linguistics)—NLP Applications track
- Area chair of COLING 2018 (The 27th International Conference on Computational Linguistics)—Information extraction tasks
- Publicity chair of ACL 2018 (The 56th Annual Meeting of the Association for Computational Linguistics)
- Area chair of ACL 2017 (The 55th Annual Meeting of the Association for Computational Linguistics)—Information Extraction and Retrieval, Question Answering, Text Mining, Document Analysis and NLP Applications track
- Co-chair of ADCS 2015 and 2016 (Australasian Document Computing Symposium).
- Co-chair of ALTA 2013 (Australasian Language Technology Association Workshop)

Steering/executive committee membership/advisory board:

- Advisory Board, Widening NLP, an ACL-affiliated affinity group dedicated to increasing the visibility and participation of researchers from underrepresented groups within the NLP, CL, and AI communities. (2025–2027)
- Director of Publicity, Association for Computational Linguistics (2023-present)

- Deputy Director of Publicity, Association for Computational Linguistics (2019-2022)
- Australasian Language Technology Association (2015-2022)
- Australasian Document Computing Symposium (2016-2023)

Journal Reviewer:

A complete list of reviews is available on Publons.

- Artificial Intelligence in Medicine (AIIM)
- BMC Medical Informatics and Decision Making
- Journal of Biomedical Informatics (JBI)
- Information Processing and Management (IPM)
- Journal of Information Sciences (INS)
- ACM Computing Surveys
- Journal of Language Resources and Evaluation (LRE)
- Transactions of the Association for Computational Linguistics (TACL), Standing reviewer team
- ACM Transactions on Asian Language Information Processing (TALIP)
- IEEE Transactions on Knowledge and Data Engineering (TKDE)
- Information Retrieval, Springer

Conference Program Committee or Reviewer:

- The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)
- ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2021)
- The 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 9th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2020)
- NLP COVID-19 Workshop, an emergency workshop at ACL 2020
- The Conference on Artificial Intelligence (AAAI 2020)
- The Annual Symposium of the American Medical Informatics Association (AMIA 2019)
- The Annual Symposium of the American Medical Informatics Association (AMIA 2018)
- The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, Demonstrations (SIGIR 2018)
- The 15th Annual Workshop of the Australasian Language Technology Association (ALTA 2017)

- The 22nd Australasian Document Computing Symposium (ADCS 2017)
- The 14th Annual Workshop of the Australasian Language Technology Association (ALTA 2016)
- The Sixth ACL Named Entities Workshop (NEWS 2016)
- The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)
- The 13th Annual Workshop of the Australasian Language Technology Association (ALTA 2015)
- The 10th International Persuasive Technology Conference (PT 2015)
- The 5th Named Entities Workshop in conjunction with ACL (NEWS 2015)*
- The 9th Conference on Information Technology in Asia (CITA 2015)
- The 37th International ACM SIGIR conference on Research & development in information retrieval (2014) Short paper, Demo, and Tutorial tracks
- The 19th Australasian Document Computing Symposium (ADCS 2014)
- The 12th Annual Workshop of the Australasian Language Technology Association (ALTA 2014)
- The 9th International Persuasive Technology Conference (PT 2014)
- The 36th International ACM SIGIR conference on Research & development in information retrieval (2013) Short paper
- The 18th ACM International Conference on Intelligent User Interfaces (IUI 2013)
- The 2nd Workshop on Advances in Text Input Methods in conjunction with COLING (WTIM 2012)
- The 4th Information Interaction in Context Symposium (Iiix), 2012.
- The Annual International ACM SIGIR Conference, posters, 2012.
- The 50th Annual Meeting of the Association for Computational Linguistics (ACL), demonstrations, 2012.
- The Named Entities Workshop (NEWS) 2012, in conjunction with ACL, 2012.
- The 5th International Joint Conference on Natural Language Processing (IJCNLP), 2011.
- The Named Entities Workshop (NEWS) 2011, in conjunction with IJCNLP, 2011.
- The 6th Asia Information Retrieval Society Conference (AIRS), 2010.
- The 6th International Conference on IT in Asia (CITA), 2009.
- The Australasian Language Technology Workshop (ALTW), 2008.
- Reviewer for The Australasian Database Conference, ADC 2008.
- Reviewer for Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2007.
- Reviewer for The Annual International ACM SIGIR Conference, poster, SIGIR 2007.

- Reviewer for String Processing and Information Retrieval, SPIRE 2006.
- Reviewer for Australasian Language Technology Workshop, ALTW 2006.
- Reviewer for The International Conference on Information Integration and Web Based Applications Services, IIWAS 2005.

Organising committee member of RMIT

Computer Science and IT Research Student Conference (CRSC), 2006.

Volunteering

Scientist in School (2012–present). In partnership with teachers in schools, teaching some core topics in computer science (e.g., algorithms, artificial intelligence) and giving presentations to encourage students to study computer science/engineering.

Diversity and women representative officer of NICTA (National ICT Australia), Victoria research laboratory.

Postdoctoral Fellow (Co-)Mentoring

2024–present, Shakila Tonni, SQL Query generation, Genomics question-answering

2023–present, Jiyu Chen, NLP for mental health

2023–present, Huichen Yang, Multimodal LLMs, Information extraction from specialised documents

2023–present, Vincent Nguyen, Climate question-answering

2023–2024, Biaoyan (Byron) Fang, Biomedical summarisation of scientific literature

2021–2023, Xiang Dai, Family history information extraction from medical records

2019–2021, Maciej Rybinski, Clinical decision support, Search for Clinical Trials

2018–2020, Aditya Joshi, Syndromic surveillance using social media

Student Supervision

2025—present, Yidong Gan, PhD Student on NLP for Clinical Coding, Sydney University (co-supervisors: Jonathan Kummerfeld and Ben Hachey)

2024—present, Tomas Turek, PhD Student on NLP in Climate Adaptation, RMIT University (co-supervisors: Jenny Zhang, Mahesh Prakash, Dharendra Singh, Vincent Nguyen)

2024—present, Chao Wang, PhD Student on Multimodal LLMs, University of Technology Sydney (co-supervisors: Lina Yao, Huichen Yang, Yi Yang)

2022—2025, Hanna Pethani, PhD Student on Clinical Natural Language Processing, Sydney University (co-supervisors: Adam Dunn, Heiko Spallek, Mike Conway, Xiang Dai)

2019—2023, Vincent Nguyen, PhD Student on Question Answering, Australian National University (co-supervisor: Zhenchang Xing, Maciej Rybinski)

2018—2023, Wenyi Tay, PhD Student on Opinion Summarisation, RMIT University (co-supervisors: Jenny Zhang and Stephen Wan)

2018—2021, Sonit Singh, PhD student on Radiology Report Generation, Macquarie University (co-supervisors: Len Hamey, Kevin Ho-Shon)

2017—2021, Xiang Dai, PhD Student on Information extraction, Sydney University (co-supervisors: Cecile Paris, Ben Hachey)

2023—2024, Harrison Bailey, CSIRO summer student, ANU University (co-supervisors: Xiang Dai, Necva Bolucu, Melanie McGrath)

2023—2024, Ritvik Dinesh, CSIRO summer student, Sydney University (co-supervisors: Xiang Dai, Byron Fang)

2023—2024, Hashem Hijazi, CSIRO summer student, Sydney University (co-supervisors: Diego Molla, Vincent Nguyen)

2022—2023, Angel Wu, CSIRO summer student, Monash University (co-supervisor: Xiang Dai)

2022—2023, Mong Yuan Sim, CSIRO summer student, Adelaide University (co-supervisor: Xiang Dai)

2022—2023, Skyler Zou, CSIRO summer student, UNSW (co-supervisor: Xiang Dai, Grant Brinkworth)

2021—2022, Liam Watts, CSIRO summer student and internship, Sydney University (co-supervisor: Maciej Rybinski)

2020—2022, Aleney Khoo, CSIRO summer student and honours thesis, University of Sydney (co-supervisor: Maciej Rybinski, Adam Dunn)

2016—2017, Xiang Dai, Industrial trainee, University of Sydney

2016—2018, Vincent Nguyen, CSIRO summer student, University of Sydney

2014—2015, Han Zhao, CSIRO summer student, University Technology of Sydney

2013—2013, Allan Jun Yin, Industrial trainee, Australian National University

2012—2013, John Lingad, CSIRO summer student and honours thesis, University of Sydney

Resources Developed

A list of software prototypes/tools and datasets that are developed in different projects under my lead/supervision (credit to many colleagues and students who were involved):

MyClimate CoPilot (2025): A conversational question-answering system for climate adaptation in agriculture.

FareBio Dataset (2024): A dataset for understanding faithfulness of LLMs on plain biomedical summaries.

MedRedQA dataset (2024): A large non-factoid English consumer Question Answering (QA) dataset.

CADECv2 dataset (2024): An extension to the CADEC dataset.

SearcEHR (2021): A family history search system on electronic health records for clinical decision support.

Science2Cure (2021): A Clinical trial search prototype.

A2A (2020): A public platform for searching biomedical literature. It allows fair benchmarking of ranking functions and document (pre-)processing algorithms..

CADEC dataset (2015): A dataset of Adverse Drug Reactions in Patient Reports. (cited 300+)