(+61) 402780774
sanjee.io
sanjeevan.selvaganapathy@uwa.edu.au

Sanjeevan Selvaganapathy

Australian Citizen
Google Scholar

Research Interests

Evaluation and alignment of large language models, with a focus on cultural pragmatics, ideological robustness, and the gap between model knowledge and behavior. Broadly interested in how LLMs encode and deploy social knowledge.

Education

University of Western Australia, *Bachelor of Philosophy, Honours - Computer Science. Undergrad GPA: 6.7/7* 2020 — 2024
Thesis: Beyond Words: Harnessing Large Language Models for Detecting Implicit Hate Speech (Higher Distinction/First Class Honours). Supervised by Dr Mehwish Nasim.
University of Nottingham UK, *Exchange - Computer Science* 2022 — 2023

Publications

Confident, Calibrated, or Complicit: Probing the Trade-offs between Safety Alignment and Ideological Bias in Language Models
S. Selvaganapathy, M. Nasim Under review, targeting ACL 2026 (ARR meta-review: 4/5)
arXiv:2509.00673
Demonstrated that safety-aligned LLMs resist ideological manipulation significantly better than uncensored models (78.7% vs 64.1%), while identifying calibration failures that create fairness disparities in content moderation.

Activation-Space Personality Steering: Hybrid Layer Selection for Stable Trait Control in LLMs
P. Bhandari, N. Fay, **S. Selvaganapathy**, A. Datta, U. Naseem, M. Nasim EACL 2026
arXiv:2511.03738
Developed activation steering methods to reliably control Big Five personality traits in LLM outputs using hybrid layer selection.

Fifteen Percent Fluency: Measuring the Cultural Knowledge-Behaviour Gap in LLMs
Authors anonymized for review Under review, ARR January 2026
arXiv
Introduced a metric to quantify how well LLMs apply cultural knowledge in practice; found models deploy only ≈15% of their demonstrated cultural capability in naturalistic contexts.

Research Experience

Research Assistant – University of Western Australia **April 2025 –**
- Developed a stochastic multi-agent simulation engine modeling social discourse dynamics, using LLMs initialized with Big-Five personality profiles to study opinion propagation and emergent polarization.
- Implemented dynamic network topologies (Erdős-Rényi, Barabási-Albert) to benchmark sentiment diffusion across varying connectivity and agent homophily conditions.
- Built and deployed a Retrieval-Augmented Generation (RAG) system for confidential research data, including full-stack implementation (Next.js, Vercel, PostgreSQL) and custom ingestion pipelines compliant with Australian data legislation.

Teaching Experience

Teaching Assistant, Relational Database Management Systems – University of Western Australia **February 2025 –**
- Delivered workshops to 50+ students on advanced database topics; facilitated weekly lab sessions.
- Graded assignments and exams; provided individualized tutoring on complex material.

Industry Experience

Software Engineer – Enaccess Maps **March 2024 – February 2025**
- Designed and built a full-stack accessibility mapping platform (Next.js, TypeScript, PostgreSQL) for a nonprofit startup.
- Implemented caching and optimization systems, reducing query latency by 60%.
- Built WCAG-compliant interfaces with secure authentication flows.

Skills

Programming	Python, TypeScript, SQL
ML/NLP Tools	PyTorch, HuggingFace Transformers, Ollama, OpenAI API, scikit-learn, spaCy, NLTK
Data & Visualization	NumPy, SciPy, Pandas, Matplotlib, Seaborn
Research Methods	Experimental Design, Statistical Hypothesis Testing, LLM Evaluation & Benchmarking, Prompt Engineering, Dataset Curation, Activation Analysis
Other	LaTeX, Git, React, Next.js