Gareth Roberts

Principal AI Specialist | Production AI, Agentic Systems, Evaluation and Infrastructure

Sydney, Australia
gareth.roberts@ieee.org
+61 499 773 667
linkedin.com/in/dr-gareth-roberts
github.com/dr-gareth-roberts | 2.works

Founding / staff-level AI systems leader with 15+ years delivering production ML and agentic AI systems across HR tech, insurance, mining, and health. Track record building AI functions from zero, shipping multi-agent systems at scale, designing evaluation and validation loops, and leading heavy-compute ML platforms. PhD in Cognitive Neuroscience; comfortable across TypeScript, Python, SQL, product judgement, and production operations.

25M	4.4x	<24h
Users	**Throughput**	**Compute**
Production multi-agent coaching systems at enterprise scale.	Regulated AI underwriting system built from zero.	Geospatial ML pipeline reduced from 6+ weeks.

Experience

Principal AI Specialist | Culture Amp

2025 - present

- Architected production multi-agent LLM systems using LangGraph and CrewAI for autonomous multilingual coaching across 20+ languages, serving 6,500 enterprise clients and 25M employees.

- Built evaluation frameworks with multi-agent translation validation to measure accuracy, fairness, and behavioural reliability.

- Engineered deterministic state guarantees and audit trails for bias-detection and fairness pipelines.

Head of Artificial Intelligence | NEOS Insurance Group

2023 - 2024

- Built the AI function from zero; recruited 6 ML engineers / data scientists, set technical direction, and delivered production AI in a regulated insurance environment.

- Designed a multi-agent RAG underwriting system across structured and unstructured inputs; achieved 4.4x throughput improvement and materially reduced quote-to-bind cycle time.

- Implemented model validation, monitoring, and audit procedures aligned to APRA-grade operational expectations.

Chief Technology Officer | Source Localisation

2022 - 2023

- Built production computer vision and geospatial ML systems for Tier-1 mining clients and scaled the engineering team to 12.

- Developed deep learning pipelines that collapsed exploration target identification from 6+ weeks to under 24 hours.

- Designed terabyte-scale data pipelines for feature extraction, dimensionality reduction, and predictive modelling.

Selected consulting and contract work | Bluesana, Comtrac, BrainChip and Bonsai Tech

2008 - 2025

- Built AI matching, RAG, and vector-database systems for sustainability analytics and enterprise document workflows.

- Designed intelligent document processing architecture for legal and claims workflows using LLM-powered classification, extraction, and semantic analysis.

- Published on LLM misbehaviour, prompt injection, and cross-model behavioural comparison; 480+ citations and h-index 8 across research outputs.

- Founded and led a data science consultancy delivering end-to-end ML across pharma, mining, education, and government.

Selected additional engineering and research

Academic and applied work spanning neuroscience, signal processing, and AI systems. Co-founded research groups, supervised PhD candidates, and built large-scale analysis pipelines for EEG, fMRI, and fNIRS data. This background is where my evaluation rigour comes from.

Recent writing and research focus on agent behaviour, prompt injection, and the gap between raw capability and reliable self-assessment in production systems.

Technical stack

Languages: TypeScript, Python, SQL

AI / ML: LangGraph, CrewAI, RAG, multi-agent systems, model evaluation, prompt engineering, PyTorch, TensorFlow, HuggingFace

Infrastructure: AWS, GCP, Azure, PostgreSQL, vector DBs, Docker, FastAPI, high-performance computing

Methods: experimental design, A/B testing, causal inference, psychometrics, time-series analysis, statistical modelling

Education and selected writing

PhD, Cognitive Neuroscience and Quantitative Methods - University of Western Australia

MBA - Power Business School

BA (Hons), First Class, Psychology - University of Western Australia

Technical writing and research across AI safety, prompt injection, agent evaluation, and neuroscience.

Selected outputs include work on model misbehaviour, prompt injection principles, and quantitative approaches to agent verification.