

William Catt

+61 404 027 684 | wcatt98@gmail.com | williamcatt.dev
linkedin.com/in/william-catt | github.com/WillCatt

PROFESSIONAL SUMMARY

Data Scientist experienced in insurance analytics, statistical modelling, NLP, geospatial analysis, and generative AI. Builds production-oriented Python and SQL pipelines, Azure-based RAG systems, evaluation workflows, and stakeholder-facing analytics.

EXPERIENCE

Data Scientist <i>Allianz</i>	Melbourne, Australia 2024 - Apr 2026
- Delivered reporting and applied analytics across 11 agency portfolios, maintaining dashboards, assuring data quality, and translating performance results into recommendations for product managers. - Built Python NLP pipelines using spaCy and regex to map noisy free-text product categories into a structured taxonomy, producing consistent classifications for downstream analysis. - Engineered geospatial features from shapefiles and Australian Bureau of Statistics Mesh Block data to segment portfolio performance, enriching datasets through geocoding API integrations. Key Project: QuoteGuard - Guardrailed RAG Chatbot - Built a grounded RAG assistant for Australian insurance Product Disclosure Statements using Azure OpenAI and Azure AI Search, combining hybrid retrieval, cross-encoder reranking, and grounding checks to achieve recall@5 of 0.80. - Designed guardrails for financial advice, pricing, and prompt-injection attempts, improving the adversarial safe-rate from 0.73 to 1.00; evaluated with RAGAS and tracked experiments in MLflow.	

ADDITIONAL EXPERIENCE

Support and Warehouse Manager RAMAWORKS	2021 - 2022
Metal Analyst BVAQ	2020 - 2021

SELECTED PROJECTS

Legal Text Anonymiser <i>Python, RoBERTa, Hugging Face Transformers, spaCy</i>	2026
- Fine-tuned a RoBERTa NER model on legal text to detect and redact PII, reaching F1 0.85 across entity types; combined the transformer with spaCy recognisers and regex rules tuned for Australian formats. - Implemented mosaic re-identification risk scoring to quantify residual disclosure risk, plus reversible pseudonymisation that consistently restores entities for authorised users.	
Touchline - Agentic Football Analyst <i>Python, SQL, SQLite, Claude, Ollama</i>	2026
Built a read-only tool-use agent that converts natural-language questions into SQL over 49,363 international matches, executes and charts results, self-corrects query errors, and benchmarks frontier and local models on 30 questions across accuracy, groundedness, latency, and cost.	

TECHNICAL SKILLS

Cloud & ML Platforms: Azure Machine Learning, Microsoft Fabric (ML), Azure Databricks, MLflow, Delta Lake, AWS (Lambda, API Gateway)
Languages & Libraries: Python, PySpark, SQL (PostgreSQL, T-SQL, SQLite), R, JavaScript, HTML/CSS; React, scikit-learn, PyTorch, Keras, XGBoost, spaCy, NLTK, pandas, NumPy, SciPy, statsmodels, NetworkX, lifelines
GenAI / LLM: Azure AI Foundry, LangGraph, LangChain, PydanticAI, RAG, embeddings, agentic workflows, tool-use agents, Ollama
MLOps & DevOps: Git, GitHub Actions, Azure DevOps, CI/CD, Bicep (IaC), Docker, pytest, ML evaluation
Tools: R Shiny, D3.js, Tableau, Streamlit, Gradio, Jupyter
Certification: Microsoft Certified: Azure Data Scientist Associate (DP-100)

EDUCATION

Monash University <i>Master of Data Science</i>	Melbourne, Australia 2022 - 2024
80+ WAM (High Distinction average); Sir John Monash Scholarship for Academic Excellence - Recipient, 2025 Cliff Bellamy Award for Best Postgraduate Industry Experience Project	
Swinburne University of Technology <i>Bachelor of Science, Chemistry</i>	Melbourne, Australia 2015 - 2019