# PATRICK S. KEOUGH

Sydney, NSW • 0403 792 091 • pskeough@gmail.com
linkedin.com/in/patrick-keough-b5b7a7261 • github.com/pskeough • Google Scholar • pskeough.github.io

AI evaluation researcher and builder. I find out what language models actually do at scale and ship the evidence: nine public studies, 140,000+ scored observations, two arXiv preprints, each released with its pipeline code and a receipt for every claim. I also build the systems: an open-source voice-cloning MCP server (11,000 lines, 153 tests), a 23,000-line retrieval memory system, and contracted software for two Sydney clients.

## Experience

**Independent AI Evaluation Researcher**, Sydney — Jun 2025 to present

- Designed and ran nine audits of language models totalling 140,000+ scored observations; the largest ran 28,800 simulated patients across four model families and 120 demographic cohorts against survey-weighted NHANES anchors.
- Built every LLM-as-judge pipeline end to end: multi-axis severity rubrics, cross-model judge validation, position and self-preference bias controls, at scale over OpenRouter and native provider APIs.
- Validated automated scoring against 236 human annotations (Cohen's $\kappa = 0.78$ with the judge) and ran a 186-claim verification pass on a co-authored manuscript before submission.
- Statistics in Python: mixed-effects models, non-parametric tests, effect sizes, inter-rater reliability, BH-FDR and Holm correction. Two preprints on arXiv, one co-authored paper with ACU in preparation.

**AI Systems Contractor**, Advanced Human Technologies (AHT Group), Sydney — Jun 2026 to present

- Built and delivered a voice-emulation pipeline for a client corpus: LLM-as-judge rubric, a measured quality gate every draft must clear, shipped as an MCP server with hybrid retrieval. Contracted and invoiced.

**Technical Consultant**, McLennan leadership diagnostics, Sydney — 2026

- Built a company-evaluation web app (5,500 lines of Python, Vite front end) that scrapes public data, scores it against three published leadership frameworks with a source-verification pass, and renders an editable executive briefing with PDF export.
- Built a Gmail classification and routing system in Apps Script for the same client: 77 unit tests and 37 integration tests against an in-memory Gmail, with structural guarantees that it cannot send, filter, or delete.

**Teaching Assistant, Statistics in Psychology (PSYCH 315)**, University of Washington — Winter 2025

## Systems

**Mimesis** (open source, github.com/pskeough/Mimesis_Voice_Clone). Stylometric voice-cloning MCP server, 11,231 lines of Python, 153 tests. Selects drafts on a 13-feature author fingerprint instead of an LLM judge; its cadence feature set separates real prose from rhythm-destroyed prose at AUROC 1.000 where an order-invariant set performs at chance.

**Claude Mind.** Local memory and retrieval system for an agentic coding tool, 23,076 lines of TypeScript: SQLite with sqlite-vec, bi-encoder recall, cross-encoder rerank, Personalized PageRank graph channel, 17 MCP tools, 2,475 captured sessions. Audited end to end: 0 scope leaks on 58 probes, p50 retrieval 3.7 s.

**Signal.** Nightly multi-agent pipeline that fans out five collector agents, dedupes against a persistent index, ranks against a feedback-tuned profile, and publishes an offline PWA; 31 consecutive nightly runs.

## Selected research

**The Granularity Gap.** arXiv:2606.05183, v3. Severity audit of 8,830 responses across three Gemini generations, judge validated on 236 human annotations. A binary refuse-or-comply verdict explains 29% of the variance in graded severity, so pass/fail and graded metrics measure different things.

**Plausible Patients, Impossible Populations.** arXiv:2604.17359, v2. 28,800 simulated assessments, four models, 120 cohorts. Every benchmarkable group is inflated against NHANES while 97.3% of single cases pass a DSM-5 gateway rule, so case-level review cannot detect the population-level failure.

Seven further public studies, including **A Validity Ladder for LLM-Simulated Clinical Populations** (arXiv, submitted Sept 2026). Code and data at github.com/pskeough.

## Skills

**Evaluation and statistics.** LLM-as-judge design, severity rubrics, cross-model validation, bias controls, human validation and IRR, mixed-effects models, effect sizes, multiple-comparison correction, survey-weighted analysis, red-teaming. **ML and LLM.** Large-N inference pipelines, RAG and hybrid retrieval, embeddings and rerankers, LoRA/SFT/DPO fine-tuning and adapter serving, multi-GPU and rented-GPU training, abliteration, SAE interpretability, stylometry, local inference (llama.cpp), MCP servers, agentic workflows. **Engineering.** Python (pandas, NumPy, SciPy, statsmodels), TypeScript, React, Flask, SQLite, pytest, Git, LaTeX.

## Education and leadership

**BA Psychology, University of Washington.** Conferred June 2025. GPA 3.77/4.00. Dean's List every quarter.

**Students for Sensible Drug Policy.** National Executive Policy Board, one of two student representatives (2022 to 2023). Founded the first naloxone distribution station at UW; ran overdose-response trainings with the Washington Department of Health.