PATRICK S. KEOUGH

Sydney, NSW, Australia
pskeough@gmail.com • github.com/pskeough • pskeough.github.io

Research statement

Independent AI-evaluation researcher working at the intersection of psychological measurement and language-model behaviour. My work turns specific failure modes, sycophancy, demographic bias and unfaithful explanation, into large-N empirical audits. Nine public studies, two live on arXiv and a third in submission, 140,000+ scored observations, each with a public repository of pipeline code and claim-level receipts. The methodological commitment is measurement adequacy: continuous severity rubrics over binary verdicts, judges validated across models and across labs, and estimators checked against reference implementations before use. A second line of work applies the same standard to creative writing: custom per-author models and the training behind them, the tools writers use to work with them, and a provenance layer over what may be used for training. I write fiction and essays, and every corpus involved is my own.

Preprints

The Granularity Gap: A Multi-Dimensional Cross-Generational Audit of Sycophancy in Gemini Models. arXiv:2606.05183, 2026; v3 live 28 Aug 2026. The v2 revision withdrew three v1 claims and added a four-judge panel. Target: FAccT / AIES. github.com/pskeough/The-Granularity-Gap.
Sycophancy-severity audit across three Gemini generations, N=8,830 responses from 8 variants, judge validated against 236 human annotations (Cohen's κ=0.78). The judge's own refuse-or-comply verdict explains 29% of the variance in its own continuous severity scores and no function of that axis explains more than 35%, so the binary metric and the graded one are close to different measurements. Released with 10,792 per-vote judge scores and written reasoning.

Plausible Patients, Impossible Populations: Auditing Epidemiological Fidelity in Large Language Model Mental Health Simulations. arXiv:2604.17359, 2026 (v1 as *PsychBench*; v2 live 5 Aug 2026 with author-rederived NHANES ground truth and a published changelog withdrawing four v1 findings). github.com/pskeough/plausible-patients.
Epidemiological audit of 28,800 simulated assessments across 4 models and 120 demographic cohorts. Every benchmarkable group returns inflated by 2.8–5.5 PHQ-8 points against survey-weighted NHANES anchors, while 97.3% of individual cases satisfy a formal DSM-5 gateway rule against a null of 10.4%: single-case inspection cannot detect the population failure. Every anchor re-derived from primary microdata and validated against published targets before use.

A Validity Ladder for LLM-Simulated Clinical Populations. arXiv preprint, submitted September 2026. github.com/pskeough/A-Validity-Ladder-for-LLM-Simulated-Clinical-Populations.
A validation protocol for a generated patient population, built because case review is where validation usually stops. A regeneration-stability gate sits beneath four levels, individual coherence, subgroup fidelity, population calibration and structural fidelity; each targets a failure the levels below it cannot detect, and passing a level licenses a stated use rather than general fitness. The ladder is read per model and records a stop wherever no reference exists. Demonstrated on the 28,800-assessment corpus against survey-weighted NHANES anchors: a single prompt changes severity category on a third of its repeated draws, cases pass the coherence rule in three of four models, and demographic gaps survive in at most one.

Working papers

Same Bias, Different Mechanisms: Socioeconomic Cues in LLM Mental-Health Assessment. 2026. github.com/pskeough/psych_scope.
Counterfactual audit of socioeconomic-cue bias across three model families, paired with a sparse-autoencoder test of whether interpretability accounts for the effect it measures.

Telling More Than They Can Know: Verbal Reports on Model Processes in Clinical Scoring. 2026. github.com/pskeough/telling-more-than-they-can-know.
Explanation faithfulness across four models; 1,430 scores, 5,594 explanations, three-vote judge consensus. Socioeconomic status drives 48.75% of score variance and receives 17.2% of the models' own feature-attribution weight.

Access, Not Capability: What Retrieval Buys Small Local Models on Private-Corpus QA. With Professor Robert H. Tai (Australian Catholic University). TMLR target. github.com/pskeough/access-not-capability.
Closed-book frontier models lose to an 8B local model with retrieval in 95.8% of comparisons; given identical retrieved context a grounded judge cannot separate them in 77.4% of pairs. Judged blind in both presentation orders; 186-claim adversarial revalidation shipped with the artifact. The private corpus, whose reference answers were model-generated while the judge shared their model family, now carries an independent expert human gold of 50 items authored without sight of the machine gold, closing the one unaddressed threat to the headline comparison.

Deeper Than the Guardrails: Demographic Bias Survives Refusal-Direction Removal in Standardized Psychiatric Screening. 2026. github.com/pskeough/deeper-than-the-guardrails.
Ten base/abliterated model pairs, 95,920 observations. Over-pathologization falls 2.13 PHQ-8 points under abliteration (21 of 21 comparisons) and still sits roughly 9 points above population norms, so the bias is not held in place by

the refusal mechanism.
Two further public studies sit outside this line of work: image preprocessing for vision-model transcription (github.com/pskeough/leave-the-image-alone) and semantic-to-parametric audio mapping graded against working engineers (github.com/pskeough/the-listening-gap).

In preparation, with Professor Robert H. Tai: a two-axis reliability methodology for LLM-assisted qualitative coding, and a reliability-validity dissociation on real interview data (IJQM target).

In progress: sycophancy under abliteration (preregistered pilot).

What an Adapter Buys a Per-Author Model, and What It Costs. Manuscript in preparation, 2026.
A measurement study of per-author adaptation across 13 authors, two model families (Qwen3 and Ministral 3) at four parameter counts, 110 runs and 40,726 generated passages against 546 held-out briefs. A LoRA adapter moves a model toward the author on embedding instruments and away from the brief on construction and adherence, at every parameter count, and model size changes neither effect; corpus size is the moderator that survives, the adherence cost shrinking past roughly 50 to 110 training briefs. A 4B adapter matches three hosted frontier models on the voice instruments above about a hundred briefs and loses below fifty, while the hosted models hold the brief 44 to 45 coverage points ahead and sit within a replicate floor of one another, so the hosted price tier buys no voice. Every contrast is read against the replicate floor of its own measure, the largest difference between two generations of one run at identical settings, so a contrast under its floor is within noise whatever its p-value.

Research systems

All built and maintained solo and run locally without cloud API keys. The public repositories ship with no author or client data.

Mimesis. *Stylometric voice cloning with a measured gate. github.com/pskeough/Mimesis_Voice_Clone*
An MCP server and CLI (∼3,600 lines, 13 modules) built on the premise that prompt-only imitation fails stylometric verification. Rather than trusting an LLM judge, it generates a slate of candidates and selects on distance to a 13-feature stylometric fingerprint calibrated against a leave-one-out corpus self-baseline. Style anchors come from reciprocal rank fusion over dense and lexical indexes, then maximal marginal relevance, so the model sees a spread of the author rather than paraphrases of one passage. Three gates follow: per-author surface rules, a source-versus-output fidelity audit on numbers and citations, and a detector signal reported but never used to force the mean, which collapses the voice toward the corpus average. A held-out discrimination harness with leak controls evaluates the result, and generated output falls inside the corpus's own self-baseline.

Voice-panel reliability study. *An inter-rater analysis that went against the system it evaluates.*
The evaluation layer over the voice work, run across 106 ground-truthed discrimination trials. Every estimator is implemented from its definition against numpy alone and validated before use: a 19-check suite pins Krippendorff's α to a reference implementation across 200 random matrices with missing data at $<$1e-10 absolute error, Fleiss' κ to statsmodels, and ICC(2,1) and ICC(2,k) to Shrout and Fleiss (1979). The three-model judge panel returns $\alpha = -0.183$ [-0.247, -0.134], systematic disagreement rather than noise, because one judge is inverted (0 of 22 on a two-alternative task, exact binomial p=4.8e-07, no position preference). Voice was the least reliable of seven rated axes, ordinal $\alpha = -0.163$ against $+0.328$ for judged insight, and the automated stylometric measure separates real from generated at AUROC 0.731 while ranking generated arms at 0.458, which is chance.

Claude Mind (LKHS). *Local hybrid-retrieval memory for an agent, over a personal vault.*
A TypeScript system pairing a dense substrate with a compiled graph index. Chunks carry document context before embedding, and retrieval runs bi-encoder recall over SQLite with a vector extension then cross-encoder reranking, behind calibrated inject and skip thresholds. Memory is layered from chunks up through session journals and per-project state cards to cross-project theme cards, with a separately weighted identity layer of roughly 1,264 dated facts carrying provenance and a clinical tier quarantined from automatic injection. A Louvain-community graph build, retrieval-quality eval harnesses and calibration scripts sit alongside it. All embedding and reranking is local.

Epithelium. *Provenance and consent enforcement over a personal literary corpus.*
A graph of 566 nodes and 3,021 edges across 181 of my own works, in which every work carries `voice_safe` and `contamination` flags in its frontmatter and the query layer refuses to return a work for style purposes unless both pass. The ruling is split by use rather than by file, so a text may inform plot, theme and character while being barred from informing diction and rhythm. 175 works clear the provenance gate and 27 form the curated style set. The screen behind it scores 12 register and orthography features on robust z against a 144-work baseline, flagging at three features past $|z| = 2$. A positive control found it blind to the case it was built for, since voiced prose passes by undershooting the corpus rather than exceeding it, and that result moved the generator's own gate to a two-sided test.

Claude CoWrite. *A local-first manuscript editor built around word-level diffs.*
React and TypeScript over a Node service running a headless model as a subprocess, streaming to a three-pane interface: reference material, working manuscript, model. Three modes separate read-only analysis, context generation and editing. Edits return as word-level rather than line diffs, because prose changes turn on word order, grouped into hunks the writer accepts or rejects individually against a pre-edit snapshot. Flat files on disk, no database, nothing uploaded. In working use; no formal evaluation has been run against it.

Methods

- **Evaluation design.** LLM-as-judge architecture, multi-axis rubrics, continuous severity scoring, best-of-N consensus, cross-model and cross-lab judge validation, blind and order-controlled presentation, adversarial prompting.
- **Statistics.** Non-parametric methods, mixed-effects models, effect sizes, BH-FDR correction, bootstrap and exact tests, inter-rater reliability (Krippendorff's α, Fleiss' κ, ICC), power analysis, cluster and author-level bootstrap, test-retest floors as a significance precondition, survey-weighted reference anchors from primary microdata.
- **Infrastructure and domain.** Large-N evaluation pipelines; local fine-tuning (SFT, DPO, LoRA) and adapter serving; cross-family replication on rented GPUs with hash-verified provenance; abliteration; sparse-autoencoder interpretability; hybrid dense and lexical retrieval; stylometry and authorship attribution instruments. Psychology research methods, clinical instruments, explainable AI, creative writing.

Education

B.A. Psychology, University of Washington, Seattle June 2025

GPA 3.77/4.00 (180 credits earned). Dean's List. Undergraduate Teaching Assistant, PSYCH 315, Understanding Statistics in Psychology.

Service and leadership

Students for Sensible Drug Policy — Executive Policy Board (national) 2022–2023

One of two national student representatives, voting on state and federal harm-reduction policy proposals for the national membership.

Students for Sensible Drug Policy, University of Washington chapter lead 2022–2024

Founded UW's first naloxone distribution station; ran overdose-response trainings with the state Department of Health.

Writing and distinction

Transubstantiation, with an accompanying process note, submitted to *Ensemble Park* issue 3, August 2026. **The Basilisk**, a literary post-apocalyptic novel in revision, fullest draft ∼71,000 words. **The Second Tree**, 31 philosophical travel essays, 2026. Roughly 300,000 words of provenance-screened fiction and essays, written without AI. USA Swimming, B-Final, 100m Butterfly, U.S. Olympic Trials.