# CARA JADE CATALANO

cara@carajadecatalano • carajadecatalano.com

## PROFESSIONAL SUMMARY

Eight years building measurement rubrics from scratch — gravitational-wave glitch classification, 10-layer acoustic training data architecture, BCI intent detection for a non-verbal subject — each demanding novel benchmark design built from first principles. LIGO's Gravity Spy: 17,578 classifications, a 342-node frequency-stratified taxonomy adopted as the community standard, a Syracuse Undergraduate Research Fellowship, and a co-authorship offer on the Nobel Prize-winning black hole collision paper, turned down to pursue this work independently. Psychometrics and experimental psychology are the through-line: inter-annotator agreement design, cut-score derivation, longitudinal regression testing, distributional shift detection, annotation quality at scale — the toolkit this built maps onto the core question for Training Insights: tracking how model capabilities emerge and shift across training runs. Three published papers with DOIs (OSF Preprints, Zenodo); more than a dozen more in pipeline.

## WHAT DRIVES THIS WORK

My older brother Abe has lived with severe hypoxic-anoxic brain injury since a near-drowning accident in 1981 — before I could walk. He can't walk, is G-tube fed, and has been functionally non-verbal his entire life, communicating through microgestures, gaze shifts, and behavioral sequences that standard NLP pipelines cannot parse. The Abe BCI dataset — cross-modal proximity analysis with Poisson-normalized onset density and bootstrapped confidence intervals — exists because I am building a model that can read what he is trying to say. The underlying research question is not “what did this person say” but “what is this person trying to say.” Intent-recognition AI for non-verbal and limited-verbal populations is an unbuilt product category. This research, at Anthropic's scale, is the closest path to building it.

## EXPERIENCE

**Quality Control & Systems Specialist** | HeyTutor | May 2023 – 2026

- **Evaluation framework:** 16-category rubric with severity tiering and LLM/regex routing across 2,000+ education PDFs: deterministic → regex; subjective → lightweight model; high-stakes → full Claude — cost/accuracy tradeoff encoded into routing architecture. NEEDS_HUMAN escape-valve routes to human review when automated confidence is insufficient, preventing false findings from reaching tutoring decisions for thousands of students.
- **Quality methodology:** cut-score thresholds from empirical distributional analysis — standardized psychometric practice; explicit construct-boundary disambiguation rules preventing rater drift across overlapping categories (inter-rater reliability applied to LLM-as-rater); failure-mode analysis and distributional shift monitoring built from scratch for a domain with no prior measurement standard.
- **Curriculum infrastructure:** serverless internal tool for 12 users — S3 + CloudFront → API Gateway → Lambda → DynamoDB — with WebSocket real-time state sync and JWT-verified Google Sign-In; structured extraction pipeline converting Google Docs/Sheets to queryable JSON (methodology and architecture IP; content retained by HeyTutor).
- **Pipeline governance:** phase-gated review with approve/abort controls, structured audit trails, and a proposal → review → approval → deploy governance model making every evaluation decision visible and transferable.

**Independent Researcher & Systems Developer** | Self-Directed | 2016 – Present

***Voice Research — eval infrastructure for a generative speech model (2024 – Present)***

- **Eval instrument audit:** 44 custom non-voice detectors; caught and fixed a merge-loop bug that silently corrupted every detector confidence score to 1.0 — a silent data corruption bug that would have invalidated all downstream training metrics without any error signal; surfaced through threshold-parameter audit.

- **Measurement pipeline:** 162 source recordings → 160 analysis-ready clips → 3,326 XTTS (text-to-speech fine-tuning model) training clips; 83-label annotation schema distinguishing recording-level labels (stable across a clip) from segment-level labels (re-measured per segment) — preventing source-constant variables from being re-answered as noise; dedicated involuntary-event labels (swallow, throat-clear) separated from intentional vocal content.
- **Two-objective design:** kept environmental variability intact in the acoustic-analysis corpus — noise IS signal for measurement — while denoising only the TTS training corpus; defended the split against uniform-preprocessing pressure; separate measurement objectives, separate pipelines.
- **Baseline & crossover:** replaced time-of-day proxy with formal physiological; within-subject crossover (A/B testing); when construction noise invalidated a session, re-anchored to a shared pre-substance baseline — confound management in real time.
- **Controlled variable experiment:** held average pitch stable while pitch variability collapsed 438 → 288 Hz — single-variable isolation identifying variance reduction as the operative acoustic effect; experimental psychology methodology applied to ML training data generation. The goal was causal isolation: What is causing this effect, and which single variable can I hold responsible?
- **Training data quality experiment:** v3→v4 controlled test isolating data-noise from model-capacity as the loss-floor bottleneck — the same question Training Insights asks at scale: Which training variables drive capability change? Failure decomposed into independently addressable dimensions: accent bleed (data provenance) and audio fidelity (corpus preprocessing).
- **Switch-access interface for Abe (published):** continuous-output BCI design using sustained engagement as a motor signature for a non-verbal subject; Poisson-normalized chance correction (statistical test for non-random communication) across modalities. OSF Preprints, DOI 10.31224/6892.

***AI Agent Infrastructure & Agentic Evaluation***

- **Multi-agent evaluation infrastructure:** Claude-native agent federation spanning 21 project domains — custom orchestration protocols, tool-use chaining, skill/slash-command authoring, and cross-session JSONL messaging with separated planning and execution contexts; SSH-linked distributed compute (primary laptop + GPU desktop for benchmark execution and batch annotation) — structurally parallel to preference-data collection and reward-model training pipeline separation.
- **Training data provenance:** per-domain ChromaDB vector stores with semantic indexing; externally-sourced data quarantined in a decontamination pipeline before promotion to knowledge graph. Production-grade training data quality control, already running.
- **Cross-domain capability synthesis:** health metric aggregating evaluation signals from 21 parallel project domains — explicit scope definitions, exclusion criteria, and rolling activity windows — designed to surface how capability patterns emerge and diverge across domains; the same synthesis model Training Insights applies across training runs. Built the exact model Training Insights needs.
- **Capability-leakage incident:** detected, quantified, and remediated 261 non-functional stub agents that accumulated silently over five days; diagnosed root cause via SSH; verified cleanup via cross-session audit — a complete detect-quantify-remediate-verify loop on a real pipeline failure, documented as a formal incident.
- **Long-horizon agentic evaluation:** overnight 8-hour unattended evaluation pass with 5-stage filter cascade; shadow-mode (observe-only) deployment pattern for controlled-exposure evaluation before any agent goes live; phase-gated orchestration with approve/abort controls.

***AI Safety, Red Teaming & LLM Output Evaluation (2023 – Present)***

- **Safety governance:** hard 4-criteria integrity gate, adversarial multi-agent review council with named dissent roles, pre-tool-use hard-block on unauthorized writes — safety rules load-bearing, not advisory; encoded structurally, not rhetorically.

- **Red teaming:** systematic forensic analysis of Microsoft Copilot model accountability failures (Feb 2026): capability elicitation failures and multi-turn behavioral inconsistencies documented for external review. Independent AI bug report filed June 2023 — proactive failure-mode identification before formal safety evaluation frameworks for deployed models were widely adopted.
- **LLM output evaluation pipeline:** six-agent system grading generative model outputs on image-text alignment, prosody, spec compliance, grammar, structural format, and couplet integrity; pixel-whiteness cut-score system auditing 194 findings across 15 titles.

***LIGO (Laser Interferometer Gravitational-Wave Observatory) Citizen Science & Research (2016 – 2019)***

- **Gravity Spy:** 17,578 glitch classifications to train Machine Learning systems for astrophysicists (78% of 22,550 total across 27 scientific projects at Zooniverse); 342-node taxonomy built around frequency as the primary axis — chosen because spectrograms are visually read by frequency, making the system technically correct and learnable by non-physicists; epistemological edge encoding (solid = confirmed, dashed = tentative).
- **Volunteer Moderator** (2016 – present): governed inter-rater agreement standards enabling reliable training labels across thousands of citizen scientists for LIGO-Virgo-KAGRA observing runs. Enabled reliable ML labels at Nobel-Prize-winning scale.

## PUBLICATIONS

- Catalano, C. J. (2026). "Design of a Continuous-Output Switch-Access Interface for Users Whose Structured Motor Signature Is Sustained Engagement with Modulation." OSF Preprints. DOI: 10.31224/6892
- Catalano, C. J. (2026). "Hardware-Constrained Voice AI: A Split-Compute Pipeline Architecture for Single-Speaker Speech Research on Consumer Hardware." Zenodo. DOI: 10.5281/zenodo.19900630
- Catalano, C. J. (2026). "The True Cost of Independent Voice AI Research: A Full Replication Cost Analysis from Zero." Zenodo. DOI: 10.5281/zenodo.19875772
- Additional papers in active pipeline at carajadecatalano.com (see below for login)

## EDUCATION & HONORS

**Associate of Arts** | Mt. San Antonio College | Social Sciences & Machine Learning

- 4.0 GPA • Honors • Phi Theta Kappa Honor Society • Person of Distinction Award: Personal Achievement
- Full-ride transfer invitations from Cornell University, Columbia University, and the University of Pennsylvania — I declined their generous invitations to avoid uprooting my son during his high school years.

**Undergraduate Research Fellow** | Syracuse University | Gravity Spy Program | 2016 – 2018

## SKILLS & TOOLS

**AI & ML:** Claude (primary — multi-agent orchestration, tool-use chaining, agentic pipeline design, cross-session JSONL protocols, skill/slash-command authoring, LLM routing, Claude Code CLI) • Microsoft Co-Pilot • Codex • Supergrok / Imagine • ChromaDB • XTTS • vector stores • fine-tuning evaluation

**Methods:** benchmark design • inter-annotator agreement • cut-score methodology • distributional shift detection • annotation-quality protocol design • capability elicitation • multi-turn behavioral testing • longitudinal regression testing • A/B variance reduction • red teaming • training data curation • failure mode analysis • reward model evaluation • bootstrapped confidence intervals • psychometrics • experimental psychology • controlled variable isolation

## PORTFOLIO

carajadecatalano.com — white papers, Cara Voice Lab, evaluation framework documentation, AI safety analysis, and BCI research. **Guest access:** login: guest • password: ViewPortfolio26