DENNIS J. CARROLL

AI Researcher · Mechanistic Interpretability · Open-Source ML Engineer

(845) 263-6312 · · New York, NYDennisCarrollJ@gmail.com

· · · github.com/denniscarrolljlinkedin.com/in/denniscarrolljhuggingface.co/DennyDenndennisjcarroll.com

WHAT I'M BUILDING

I reverse-engineer how neural networks actually work — not just how to use them. My independent research applies mechanistic interpretability methods (TransformerLens, activation patching, circuit analysis) to GPT-2 Small, uncovering how models store knowledge, route information, and generalize. I've developed original theoretical frameworks — the Minus One Principle and Bayesian Confidence Circuits — connecting transformer internals to first-principles mathematics. I build production-grade open-source tools, ship things to the web, and write code that tests itself. I thrive at the intersection of research depth and engineering rigor.

RESEARCH & PROJECTS

Mechanistic Interpretability Research · *Independent* · 2024 – Present

• Mapped Head 7.0 as a Name Mover router and Layer 8 MLP as distributed associative memory using TransformerLens attention pattern analysis and activation patching on the IOI task

• Documented the Hydra Effect: suppression heads compensating for ablated Name Movers — providing novel evidence of dynamic redundancy circuits in small transformers

• Developed the Minus One Principle: a unifying framework showing subtraction/difference as the atomic computational primitive across gradient descent error signals, attention circuits, and LMC branching

• Framing ongoing work as Bayesian Confidence Circuits — modeling attention head outputs as uncertainty-weighted belief propagation

• Built an interactive Mechanistic Interpretability Visualization tool for exploring head behaviors across layers

edge_dynamics · *Open Source · IoT Telemetry Compression* · 2024 – Present

Python · zstd · Per-topic adaptive dictionaries · Statistical anomaly detection

• Designed an IoT telemetry compression system with per-topic zstd dictionaries, delta compression, and backpressure-aware ingestion — achieving sustained compression ratios monitored via EMA drift detection

• Built adaptive dictionary lifecycle management with schema fingerprinting to detect semantic drift and trigger retraining, fixing a fundamental flaw in EMA-only detection approaches

• Maintained 102+ passing tests across compression, ratio monitoring, and schema fingerprint modules

MCP-SP Security Protocol · *Open Source* · 2024 – Present

Model Context Protocol · Security · Agent infrastructure

• Designed a security layer for MCP (Model Context Protocol) servers — authentication, rate limiting, and sandboxed tool execution for AI agent pipelines

• Motivated by real attack surfaces in multi-agent LLM systems; addresses prompt injection and unauthorized tool invocation vectors

Privacy-First Media Mix Modeling Toolkit · *Open Source* · *Streamlit* · 2023 – 2024

Python · Laplace DP · Bayesian inference · Streamlit

• Built a differential-privacy-enabled MMM toolkit applying Laplace mechanism noise calibrated to L1 sensitivity for marketing attribution under privacy constraints

• Shipped an interactive Streamlit dashboard enabling non-technical stakeholders to explore model outputs and budget allocation scenarios

Bayesian Airbnb Market Analysis · *Portfolio* · *Streamlit* · 2023

R · Python · Hierarchical Bayesian modeling · Live dashboard

• Built a hierarchical Bayesian model for Seattle Airbnb pricing achieving $R^2 = 0.687$ with engineered spatial and listing features

• Deployed live Streamlit dashboard with LLM-powered business intelligence layer for exploratory market queries

Browser-Based ML Training Environment · *Interactive App* · 2024

TensorFlow.js · JavaScript · Real-time visualization

• Developed Edge AI Training Lab: a real-time neural network training environment running entirely in the browser with TensorFlow.js, live loss curves, decision boundaries, and network visualization

TECHNICAL SKILLS

Interpretability: TransformerLens, Activation Patching, Circuit Analysis, Attention Pattern Visualization

ML / DL: PyTorch, Hugging Face Transformers, Scikit-learn, Keras

Languages: Python (NumPy, Pandas, SciPy), R, SQL, JavaScript

Research Tools: Jupyter, Git, Docker, VS Code, Linux (Pop!_OS)

Statistics: Bayesian inference, Differential privacy, Time series, Hypothesis testing

Visualization: Matplotlib, Seaborn, Plotly, Tableau, Streamlit

EDUCATION

M.S. Data Science · Illinois Institute of Technology · In Progress

GPA: 3.4 · Relevant coursework: Machine Learning Algorithms, Statistical Analysis, Data Mining, Big Data Analytics, Distributed Computing

B.F.A. Communication Studies · SUNY Oneonta · 2013 – 2017

Foundational training in research methods, technical writing, and applied analysis

CONTINUING EDUCATION

• Deep Learning Specialization (Andrew Ng / deeplearning.ai) — In Progress. Completed Course 1 through vectorization, broadcasting, and backprop — hand-deriving each algorithm before implementing.

• Data Science & Applied AI — Brown University School of Professional Studies (2024). Capstone: predictive maintenance model for manufacturing equipment deployed in a production pipeline.

• Writing & Editing: Word Choice and Word Order — University of Michigan (Coursera)

• Growth-Driven Design & Inbound Marketing — HubSpot Academy

• Responsible Conduct of Research — CITI Program (2023)

PROFESSIONAL EXPERIENCE

Doorman / Building Data Coordinator · 60th & Fifth Corp · New York, NY · March 2020 – Present

• Maintain and query a resident database of 40+ households — the operational data work that builds real intuition for data quality and information architecture

• Developed data-driven communication protocols between building management and residents; designed digital logs that surface actionable patterns for management decisions

Carpenter's Assistant · California Closets · Hawthorne, NY · March 2019 – March 2020

• Translated spatial measurement data into optimized installation configurations; honed tolerance for precision and iterative problem-solving

Paint Department Associate · Lowe's Home Improvement · Orangeburg, NY · May 2018 – March 2020

• Built spreadsheet systems to track department metrics and inventory reorder patterns, reducing stockouts by 15%

PRESENCE

Also builds interactive fiction universes at dennisjcarroll.com/stories

· · · GitHub (32+ repos)HuggingFace · DennyDenndennisjcarroll.comLinkedIn