Curt Tigges

ct@curttigges.com | in/curttigges | github.com/curt-tigges | San Francisco, CA

PUBLICATIONS & OPEN-SOURCE

First-Author Papers

- Language Models Linearly Represent Sentiment [arxiv] – Blackbox NLP '23
  Investigation of how LLMs build representations of sentiment.
- LLM Circuit Analyses Are Consistent Across Training and Scale [arxiv] – NeurIPS '24
  Investigation of algorithmic, component, and size stability of circuits across training and over scale.

Co-authored Papers

- SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in LLM Interpretability | [arxiv]
- Sparse Autoencoders Do Not Find Canonical Units of Analysis | [arxiv]
- Transformer-Based Models Are Not Yet Perfect At Learning to Emulate Structural Recursion | [arxiv]

Mechanistic Interpretability Tools I Have Built

- Probity: A Toolkit for Neural Network Probing [github]
- Crosslayer Coding: Cross-Layer Transcoder Training for LLMS [github]

EXPERIENCE

Decode Research **San Francisco**
Science Lead *Jul 2024–Present*

- Built first open-source library for training Anthropic-style Cross-Layer Transcoders (CLTs) on GPT-2-Small, producing a demo and public library with multi-GPU training & efficient server-based activation pipeline
- Shipped Probity, a probing toolkit now used by MATS scholars and other mech interp researchers, integrating specialized LLM versions of classic techniques as well as attention probes, k-sparse probes, and other recent innovations
- Provide ML & mech interp guidance and code for our mech interp research platform (Neuronpedia)
- Led extensive rewrite of our SAE training library (SAELens) to improve usability and add support for transcoders, crosscoders, etc.

EleutherAI Institute **San Francisco**
Research Scientist *Jan 2023–Jul 2024*

- Lead author on *LLM Circuit Analyses Are Consistent Across Training and Scale*; built pipeline to extract circuits for thousands of checkpoints on LLMs from 70M->13B and conducted analyses of structure and behavior across various dimensions
- Co-lead author for *Language Models Linearly Represent Sentiment*, demonstrating techniques for finding linear features and identifying the "summarization motif"
- Built custom path-patching/activation-patching tools and conducted experiments for models trained to perform recursion
- Trained LLMs on GPU cluster for various projects as needed, and maintained and improved the GPT-NeoX library

NCSU Ops Research & Education Lab **Raleigh, NC**
Data Scientist (part time) *Jun 2022–Nov 2022*

- Predictive demographic modelling for statewide school placement.

Taroko.io **Taipei, TW & Raleigh, NC**
Data Analyst -> Senior Data Analyst *Jun 2016–Dec 2022*

- Planned & built out data warehouse in BigQuery, integrating PostgreSQL database, Heap Analytics & company-wide data sources, providing critical ROI/behavior information needed for key product expansion decisions
- Built ML/statistical solutions for churn/revenue prediction, conversion path analysis, etc., optimizing millions of dollars of ad spend
- Developed bidding algorithms that rescued failing products, decreasing CPA by 21% and saving $100Ks of ad spend

KPIT Extended PLM **Raleigh, NC**
Software Engineer *Mar 2014–Aug 2015*

- Deployed & tested PTC Windchill customizations to product lifecycle management systems

EDUCATION

SERI MATS (Neel Nanda Interpretability Stream)	2023
Alignment Research Engineer Accelerator (ARENA)	2022
Master of Computer Science (Data Science Track) | University of Illinois Urbana-Champaign	2021
Bachelor of Science in Science, Technology and Society | NC State University	2012

SKILLS

Languages: Python (expert) | SQL (advanced) | R / C++ / PHP (working)
Packages: PyTorch | PyTorch Lightning | Transformers | Scikit-Learn | Matplotlib | Pandas | Numpy
Domains: Mechanistic Interpretability | Deep Learning | Distributed Training (DDP, FSDP, DeepSpeed) | Software Engineering