Dr Arshavir Blackwell

Los Angeles, CA
+1 510-604-3874

arshavir@me.com
LinkedIn Profile
https://arshavir.com
https://insidetheblackbox.ai
https://yourvoicecraft.ai
https://arvoinen.ai

Senior artificial intelligence and machine learning scientist and engineer specializing in natural language processing and large language models (LLMs). Working at the intersection of AI and human cognition since 1987 — from connectionist models of language acquisition to the internals of modern transformers.

Extensive experience designing adaptive, language-driven applications and directing research on how humans and machines learn and use language — applied to text mining, suggestion, and generation.

Current focus: mechanistic interpretability — what is actually happening inside large language models, and what fine-tuning changes in weight space.

Currently:

- **Founder, YourVoiceCraft** (yourvoicecraft.ai) — a privacy-first writing platform: private models trained on a writer's own archive that write in their voice, quote what they've actually written, catch drafts that contradict them, and keep their work off Big AI's servers.
- **Author, *Inside the Black Box*** (insidetheblackbox.ai) — a newsletter and companion podcast on mechanistic interpretability and artificial psycholinguistics.
- **Independent interpretability research** — LoRA fine-tuning on open models and sparse-autoencoder analysis of what fine-tuning changes at the feature level.

Experience

Arvoinen Consulting **2020–Present**
Los Angeles, CA
Principal

YourVoiceCraft (yourvoicecraft.ai):

- Designed, built, and launched a privacy-first writing platform: per-writer language models fine-tuned (LoRA) on each writer's own archive, retrieval grounded in their published work, and automatic detection of drafts that contradict what they've written.
- Full stack: model training and serving on dedicated GPU infrastructure, evaluation batteries for voice fidelity, and a production web application.

Large Language Model Prototypes:

- Built systems that use LLMs to ingest, organize, and generate content from crawled web articles — extracting sub-topics and structured information within specific subject areas.
- Designed an interface that categorizes and color-codes sub-topics, letting users navigate, select, and annotate articles of interest.
- Implemented a human-in-the-loop feedback cycle in which annotations iteratively refine LLM outputs.

CitizenNet/Condé Nast **2010–2020**
Los Angeles, CA
Chief Scientist

Targeting and Market Analysis:

- Led the development of CitizenNet, a platform that automated Facebook advertising and targeting by combining the Facebook API with machine learning.
- Applied a range of machine learning methods — deep learning, self-organizing maps, and ensemble classifiers — to optimize the platform's performance.
- Devised clustering algorithms to pinpoint demographic segments, such as "Midwest fans of One Direction who also like Electric Daisy Carnival," and classification algorithms to assess user receptivity to advertising within these segments.

Automated Article-writing Pilot:

- Built a pilot system that mined research data from social networks and, with human input, generated well-informed posts on specific subjects — integrating targeting data from the platform's social-media mining to speed research collection and improve article quality.

Fox Interactive Media (MySpace) **2008–2010**
Santa Monica, CA
Senior Research Scientist

- *Demographic Prediction Engine:* Analyzed profiles of users with known demographic values, modeled them, and predicted missing values such as age and gender for other users.
- *Buzz Tracking:* Implemented trend and rate analysis algorithms to track changes in the frequency and intensity of targeted buzz words.
- *Intent Miner:* Contributed to the development and testing of a system for extracting intents from unstructured text, such as identifying "intent to purchase a car," "intent to purchase a cellphone," "just married," and "just had a child."

Early Career

MetaLINCS, San Jose, CA | Principal Computer Scientist **2007**

- *Innovation Team:* Led development and enhancement of MetaLINCS' flagship e-discovery application.

H5 Technologies, San Francisco, CA | Senior Scientist and Director of Research **2005–2007**

- *Legal Analysis with Advanced Search:* Led research and development to improve accuracy, speed, and cost-efficiency in analyzing large document sets for specific legal cases.

Entrieva, Reston, VA | Senior Engineer and Project Lead **2003–2005**

- *Document Management Upgrades:* Directed maintenance and upgrades for unstructured document management applications, including categorization software, and led development of new solutions.

Adaptive Lava, Oakland, CA | Principal **2001–2003**

- *Peer-to-peer Artificial Intelligence:* Integrated peer-to-peer functionality with deep learning and natural language systems to make files stored on individual computers accessible across an organization or knowledge community.

Comprecorp, Nevada City, CA | Principal Scientist **2001**

- *E-mail Classifier:* Led engineering of a software engine that categorized emails and free-text documents into user-defined categories using machine learning, extensible to large-scale document management.

Ask Jeeves, Oakland, CA | Senior Engineer **1999–2000**

- *Jeeves Automation Project:* Improved accuracy and reduced costs for the Jeeves question-answering system by identifying bottlenecks in knowledge-base creation and developing adaptive automation, overseeing implementation by a dedicated team.

Education

PhD, Cognitive Science & Psychology **1995**
University of California, San Diego, La Jolla, CA

A.B., Honors Cognitive Science & Honors Creative Writing **1987**
Brown University, Providence, RI

Research Interests

Mechanistic interpretability of large language models: circuits, features, superposition, and what fine-tuning actually changes inside a network. Artificial psycholinguistics — treating LLMs as experimental subjects: what they memorize versus generalize, and how their errors compare to ours. Fine-tuning internals: what a LoRA does in weight space, and how style, knowledge, and identity are encoded. Language processing and acquisition in humans and neural networks — and, throughout, the gap between fluent output and understanding.

Publications & Patents

- Blackwell, A., Bates, E., & Fisher, D. (1996). "The time course of grammaticality judgment." *Language and Cognitive Processes*, 11(4), 337-406.
- Blackwell, A. W., & Bates, E. (1995). "Inducing agrammatic profiles in normals: Evidence for the selective vulnerability of morphology under cognitive resource limitation." *Journal of Cognitive Neuroscience*, 7, 228-257.
- Warren, W. H., Blackwell, A. W., Kurtz, K. J., Hatsopoulous, N. G., & others (1991). "On the sufficiency of the velocity field for perception of heading." *Biological Cybernetics*, 65, 311-320.
- Warren, W. H., Mestre, D. R., Blackwell, A. W., & Morris, M. W. (1991). "Perception of circular heading from optical flow." *Journal of Experimental Psychology: Human Perception and Performance*, 17, 28-43.
- Warren, W. H., Blackwell, A. W., & Morris, M. W. (1989). "Age differences in perceiving the direction of self-motion from optical flow." *Journal of Gerontology*, 44, 147-153.

- Benyamin, D., & Blackwell, A. (2013). "Systems and methods for selecting and generating targeting information for specific advertisements based upon affinity information obtained from an online social network." Patent number: US 20130218678 A1.
- Benyamin, D., Hall, M., Chu, A., & Blackwell, A. (2012). "Systems and methods for automatically generating campaigns using advertising targeting information based upon affinity information obtained from an online social network." Patent number: US 20120158518 A1.
- Benyamin, D., Chu, A., Pollock, A., Hall, M., & Blackwell, A. (2012). "Generation of advertising targeting information based upon affinity information obtained from an online social network." Patent number: US 20120158489 A1.
- Sravanapudi, A., Sutler, M., Devand, S., Kalaputapu, R., & Blackwell, A. (2007). "Analyzing content to determine context and serving relevant content based on the context." Patent number: WO 2007076080 A2.

Skills

Mechanistic Interpretability - Sparse Autoencoders - LoRA Fine-Tuning - Transformer Internals - PyTorch - HuggingFace - Advanced Artificial Intelligence (AI) - Machine Learning (ML) - Natural Language Processing (NLP) - Large Language Models (LLMs) - Computational Linguistics - Text & Text Mining - Deep Learning - Adaptive Systems Design - Text Analytics - Neural Networks - Docker - LLM APIs - Research Project Management - Human and Automated Language Acquisition - Transformer Architecture Development - Data Analysis - Software Engineering - Prototyping - Algorithm Development - User Interface Design - Knowledge Extraction and Organization - Content Generation - Data Annotation - Machine Learning Model Training and Tuning - Targeting and Market Analysis - Social Media Data Mining - Demographic Prediction - Intent Extraction - Peer-to-peer Systems Integration - E-mail Classification - Patent Development - Legal Analysis with Advanced Search