

JOSHUA FONSECA RIVERA

Berkeley, CA | josh.fonseca88@gmail.com | jfr.dev | github.com/vuciv

EDUCATION

The University of Texas at Austin

M.S., Artificial Intelligence

2026

Austin, TX

The University of Chicago

B.S., Computer Science · Dean's List (2016–2020)

2020

Chicago, IL

PUBLICATIONS

- **Steering Awareness: Detecting Activation Steering from Within.** Joshua Fonseca Rivera, David Demitri Africa. *Conference on Language Modeling (COLM)*, 2026. arXiv:2511.21399
- **Touch&Fold: A Foldable Haptic Actuator for Rendering Touch in Mixed Reality.** Shan-Yuan Teng, Pengyu Li, Romain Nith, Joshua Fonseca, Pedro Lopes. *ACM CHI*, 2021. **Best Paper Honorable Mention** (top 5%). doi:10.1145/3411764.3445099

EXPERIENCE

Anthropic Fellows Program

AI Safety Research Fellow

2026 – Present

Berkeley, CA

– Research on language model interpretability.

Independent Research

2025 – 2026

First author, Steering Awareness (COLM 2026)

– Designed and ran the study behind “Steering Awareness”: fine-tuned seven instruction-tuned models to detect and name steering vectors injected into their own activations (95.5% detection, 71.2% concept identification, zero false positives), showing detection dissociates from resistance.

Taekus (fintech)

Jan 2024 – 2026

Software Engineer → Lead Mobile Engineer

Remote

- Led a production cross-platform banking app end-to-end (Expo, React Native, TypeScript): 150+ screens across onboarding, money movement, and disputes; built a native Swift/Kotlin module and Expo config plugin for Apple/Google Wallet card provisioning.
- Cut login p95 latency from >15s to <2s by parallelizing authentication requests.
- Shipped check deposit and check mailing 0→1; rebuilt statement credits as idempotent, self-healing workflows with reconciliation, eliminating double-credit incidents and manual clean-ups.
- Scaled IP telemetry over high-risk operations, enabling fraud analytics and controls that reduced fraud cases.

Root16, Technology Consulting

Jul 2020 – Apr 2023

Software Engineer

Remote

- Led .NET/C# delivery for clients including Energizer, RSM, and Plante Moran; built a data-migration tool and Azure CI/CD pipelines; mentored junior engineers.

Human–Computer Integration Lab, University of Chicago

Jan – Jun 2020

Research Assistant (Prof. Pedro Lopes)

Chicago, IL

- Built VR/AR demonstrations (Unity, Leap Motion) and ran user studies for an exoskeleton-EMS haptic system; work contributed to Touch&Fold (CHI 2021, Best Paper Honorable Mention).

INVITED TALKS

- “Steering Awareness: Detecting Activation Steering from Within,” BlueDot Impact, 2026.

OPEN SOURCE & PROJECTS

- **Artblock** (340★) — browser extension replacing web ads with fine art from public museum and NASA collections.
- **animal-crossing-llm-mod** (319★) — LLM-driven dialogue injected into a 2001 GameCube game via live memory mapping.
- **golf.vim** (295★) / **vim-bujo** (206★) — a keystroke-golf game and a minimalist task manager for Vim.
- **Simplify AI** — founder; browser extension converting dense text to plain English using fine-tuned GPT models.

TEACHING & SKILLS

- **Teaching:** TA, Emerging Technologies, University of Chicago (Winter 2020).
- **ML/Research:** Python, PyTorch, HuggingFace, model evaluation & interpretability tooling.
- **Engineering:** TypeScript, React/React Native, Django, Postgres, AWS, C#/.NET, Swift/Kotlin.