

Arsenii Trutnev

Software Engineer · Agent Security · Crypto

Burlingame, CA · arsenii@marshall.dev · arsenii.wtf
github.com/mfinikov · x.com/thearsenii · linkedin.com/in/arseniitrutnev

Software engineer with two years of experience across agent security, machine learning reliability and smart contracts. I build sandboxes that confine a coding agent to a single folder, calibration and data-cleaning pipelines that make a classifier honest about what it does not know, and self-custody contracts where only the owner's wallet can move funds. Reporting vulnerabilities since 2024.

SELECTED PROJECTS

agentpen · a disposable machine for your AI agents

[GitHub](#) · [npm](#)

Giving a coding agent your whole computer means giving it your SSH keys, your other projects and your browser profile. agentpen gives it one folder and a container: one command opens a fresh isolated environment, and everything in it is deleted when you leave. It runs on three backends behind the same commands (Apple's container runtime, Docker and macOS Seatbelt) and picks the best one present. I measured what each backend can still reach on the network and published the table and the threat model. TypeScript, MIT.

Simple Safe · a locked box for your crypto

[GitHub](#)

Anyone can put ether in, and only your wallet can take it out. One small Solidity contract and one page with no web3 libraries: the page talks to MetaMask directly, deploys a contract your wallet owns, and builds each transaction by hand. No custody and no account. A learning project on the Sepolia test network, written to be read by beginners. Next.js, MIT.

Confidence calibration · why models lie with confidence, and how to make them honest

[Research](#) · [GitHub](#)

A softmax classifier always picks a class, even when it has no evidence. I measured the gap between stated confidence and real accuracy on a ResNet18 CIFAR-10 classifier, fitted a temperature to close it, and added a rule that abstains below a confidence threshold. Accuracy on the answers it does give went from 56.8% to over 85% without retraining the model. PyTorch, reproducible from a clean environment.

Near-duplicate detection · what duplicated data does to a model

[Research](#) · [GitHub](#)

Near-identical training images waste labelling budget and are impossible to find by hand. I extracted ResNet18 embeddings for a 50,000-image dataset, compared every pair by cosine similarity and flagged those above 0.98. Removing 176 images, 0.4% of the data, raised accuracy from 65.9% to 68.9% with the model and settings unchanged. A UMAP map makes the duplicate clusters visible. PyTorch.

EXPERIENCE

Founder · Marshall Labs

Bay Area · Apr – Aug 2026

Built security tools for AI-written code as the only engineer.

Security researcher · Independent

Remote · Feb 2024 – Aug 2025

Found and reported vulnerabilities in web hosting platforms through bug bounty programs.

EDUCATION

High school diploma · Burlingame High School

Graduating 2027