# Varun Iyer

varunviyer1@gmail.com · vaiyr.com · github.com/vaiyr · linkedin.com/in/varun-iyer

*Independent AI safety researcher and startup CTO working on **AI control** and **internals-based monitoring**. Open question: can a model be trained to pursue a hidden goal while evading its own introspective monitor?*

## Research

**Fine-Tuning Silently Breaks AI Safety Monitors** Preprint · Code · 2026

- On Qwen2.5-Coder-7B, tracked a linear shortcut probe across 5 SFT rounds: the direction **rotates 40–70° per training step**, additive steering stays clean, yet ablation shows the probe has **lost causal necessity**.
- An **internals-based monitor** with AUROC 1.0 can be causally disconnected from the feature it classifies. Proposed self-recalibrating monitors with ablation-based validation as standard practice.

**Linear Safety Probes Cannot Silence Features They Detect** Preprint · Code · 2026

- A linear probe can **detect a safety feature without actually controlling it** — suppressing its direction leaves behavior unchanged. Across 4 chat models and 2 features (refusal, sycophancy), no fixed fitting method works reliably.
- Shipped a **one-scalar calibration check** that catches these silent failures: finds a working probe when one exists, alarms when none does. Flags **17 of 17** working probes in testing; fixed methods catch 10 or 13.

**Overcoming Catastrophic Forgetting in RNNs** [Poster] Freedman Lab, UChicago · 2017–18

- Augmented recurrent networks with dendritic gating, disinhibitory pathways, and synaptic-intelligence regularization to preserve learned representations across sequential tasks.

**The Google Spectrum** [PDF] UChicago Math REU · 2017

- Derivation of Google's PageRank from Perron's theorem, spectral graph theory, and Markov-chain convergence.

## Experience

**Soma** [GitHub] 2024–present
*Co-founder & CTO — decentralized foundation-model training* *San Francisco*

- **Mechanism design for scalable oversight:** independent specialists compete on next-byte loss, exploiting train/verify compute asymmetry. Stake-weighted KNN routes targets; rewards flow to the lowest-loss weights without trusting any participant.
- Shipped the Rust protocol + Python SDK on a DAG-based consensus layer (**<0.33s finality, 100k+ TPS**). Ran the validator + fullnode fleet on Kubernetes; indexed on-chain GraphQL data into Postgres for the explorer and analytics.
- Research thread: do **embarrassingly parallel, competitively routed specialists** match monolithic training at equal active compute? (4 Qwen3-0.6B experts over Pile-clustered partitions, MMLU-Pro.)

**Glass** 2020–2023
*Co-founder & CTO — decentralized video platform* *NYC / LA*

- Built the TypeScript web interface for the platform, plus the Ethereum and Solana smart contracts and HLS-to-Arweave uploader with CDN pre-warming behind it. **Made $1M+ for video creators**; raised **$6.2M**.

**Spott** 2017–2019
*Co-founder & CEO — real-time map-based social network* *Chicago / NYC*

- Built the iOS app from prototype through UChicago's New Venture Challenge; scaled to **5,000 DAU** across the US.

## Education

**University of Chicago** 2016–2020
*B.S. Computer Science, Specialization in Human-Computer Interaction*

## Skills

**ML, Interp, & Evals** JAX, PyTorch, HuggingFace, Modal, Daytona, NNsight, Harbor, Petri
**Code & Infra** Rust, Python, TypeScript; Kubernetes, Postgres, GraphQL