Farseen Shaikh

Independent AI Safety Researcher | Mechanistic Interpretability · Sparse Autoencoders · RL Alignment
farseenshaikh20@gmail.com · India · github.com/FarseenSh · huggingface.co/Farseen0 · OpenReview

SUMMARY

Independent AI safety researcher working on mechanistic interpretability for frontier LLMs, focused on adapting sparse autoencoder features for safety-relevant tasks and on how representation choices shape alignment failures in agentic and multilingual systems. Sole-author of 4 papers across SAE adaptation, LLM evaluation, and agent representation engineering, with submissions at ACL ARR 2026 and an ACL 2026 Workshop. Team Captain on Cohere Labs' Expedition Tiny Aya 2026 (Efficiency Enigma award, awarded by Marzieh Fadaee and Tom Hosking) training Sparse Autoencoders across 61 languages on the Tiny Aya multilingual model family. Two years of applied AI engineering experience building RL agents and multi-agent LLM systems in production. Working without lab affiliation or PhD, with the goal of making frontier model internals legible, controllable, and safer.

RESEARCH INTERESTS

Mechanistic interpretability for safety · Sparse autoencoder adaptation for safety-relevant feature discovery · Multilingual safety evaluation and Global South deployment risk · RL agent alignment and reward hacking · Emergent misalignment and model organisms of misalignment · Representation engineering.

RESEARCH PUBLICATIONS

SCAR: Sparse Code Audit Retriever via Encoder-Side SAE-LoRA Adaptation

Farseen Shaikh (sole author) · Under Review, ACL ARR 2026 (March cycle)

Introduces an **encoder-side SAE-LoRA**: a novel low-rank adaptation of frozen Sparse Autoencoder *encoder* weights that preserves the decoder feature dictionary, structurally distinct from Chen et al. (ICML 2025), which adapts the language model around a frozen SAE. Bridges the gap between SAE reconstruction objectives and downstream discrimination. Achieves R@10 = 0.901 on a 232k-document corpus (vs. 0.308 for BM25) and a **37.6x improvement** over the frozen-SAE baseline, while adding only ~0.3% trainable parameters. Backbone: Qwen2.5-Coder-1.5B + JumpReLU SAE (width 16,384). All data, code, and weights open-sourced.

Can LLMs Self-Correct Table Reasoning Errors?

Farseen Shaikh (sole author) · Under Review, ACL 2026 Workshop (SURGeLLM)

First systematic cross-model study of self-correction failure modes for table reasoning across 5 frontier LLMs (Gemini 3.1 Pro, Kimi K2.5, GLM-5, Qwen 3.5+, MiniMax M2.5). Proposes **Structured Self-Correction (SSC)**, a 4-check verification chain reducing over-correction by 38-75% on TabFact. Confirms an *Accuracy-Correction Paradox*: models below ~70% base accuracy benefit from self-correction; models above are harmed by over-correction.

Do Structured Data Comprehension Skills Transfer Across Representation Types?

Farseen Shaikh (sole author) · Under Review, ACL 2026 Workshop (SURGeLLM)

Cross-format transfer benchmark of 1,724 programmatically-generated questions over identical content rendered in 5 representations (table, chart-image, chart-text, graph, time-series). Six frontier March-2026 LLMs evaluated. Finds strong transfer across text-based formats (**tetrachoric r = 0.85**) but a 20-50% accuracy gap for chart-image vs. equivalent text descriptions, indicating siloed visual structured-data comprehension.

SweeperLLM: Representation Over Training

Farseen Shaikh (sole author) · arXiv preprint, February 2026 · github.com/FarseenSh/SweeperLLM

Demonstrates that **state representation dominates training method** for LLM agents. Same Qwen2.5-14B fine-tuned with LoRA: compact ASCII Minesweeper boards yield 10-15% valid moves; an explicit frontier-constraint representation yields **100% valid moves** on all board sizes (6x6 to 30x30) without fine-tuning. Includes a 3-tier CSP solver for SFT data generation and a mechanistic diagnosis of **GRPO variance collapse** under strong SFT priors.

RESEARCH EXPERIENCE

Cohere Labs — Expedition Tiny Aya 2026 (Remote)

Team Captain, Inside Tiny Aya: Mapping Multilingual Representations with Sparse Autoencoders · Jan 2026 – Mar 2026

- Led a 3-person team training **4 Sparse Autoencoders** (BatchTopK, width 16,384, k=64) on all four Tiny Aya regional model variants (Global, Fire, Earth, Water), covering **61 languages** and ~41M tokens from a balanced CulturaX subset.
- Designed the experimental program to probe how multilingual LLMs represent 70+ languages internally: language-specific vs. universal feature decomposition, regional-variant feature redistribution, and steering of low-resource language generation.
- Mentored by **Marzieh Fadaee** (Cohere Labs, Expedition Tiny Aya 2024 captain) and **Tom Hosking** (Cohere Labs); built end-to-end evaluation pipeline (LaBSE, fastText LangID, Unicode script conformity, chrF, LLM-judge with Cohere

Command-A); released 4 SAEs on Hugging Face under Apache-2.0 at huggingface.co/Farseen0/tiny-aya-saes.

- Awarded **The Efficiency Enigma** by Cohere Labs for the project's contribution to interpretability of multilingual models. Verified credential available.

INDUSTRY EXPERIENCE

Heurist AI **(Remote)**

AI Researcher — RL Agents in Adversarial, Non-Stationary Environments · Jul 2024 – Sep 2025

- Trained reinforcement learning agents in non-stationary, adversarial environments under live distribution shift, with direct exposure to **reward hacking failure modes** relevant to alignment of deployed agentic systems.
- Built RL training infrastructure bridging LLM-based reasoning with classical policy optimization for autonomous decision-making in open-ended environments; designed reward shaping and policy stability mechanisms against adversarial actors with strong incentives to exploit specification gaps.
- Worked at the intersection of frontier LLM tool-use, agent infrastructure, and on-chain data pipelines, with a focus on agent failure modes under distribution shift.

Picke.ai **(Remote)**

AI Engineer (Intern) · Jan 2024 – Jul 2024

- Built a production multi-agent AI tutoring pipeline (tutor agent + assistant agents + content generation) deployed for end-users in education; owned the end-to-end LLM stack (prompting, retrieval, evaluation harness, model serving, front-end integration).
- Designed safety- and factuality-aware response generation appropriate for learners, with emphasis on grounded answers and pedagogical reliability.

SELECTED PROJECTS

PaperMind AI — Full-stack research-paper learning platform powered by a 9-agent AWS Bedrock pipeline

github.com/FarseenSh/PaperMind_AI · Live deployment

- 9-agent pipeline (parser → explainer → diagram gen → formula renderer → RAG indexer → chatbot → code planner / analyzer / coder) over Kimi K2.5, GLM-OCR, GLM 4.7 Flash, DeepSeek V3.2, MiniMax M2.1, Cohere Embed v4 via AWS Bedrock.
- Stack: React 19 / TypeScript / Vite, FastAPI / asyncpg, RDS PostgreSQL + pgvector, Cognito OAuth, S3, Step Functions, EventBridge, CloudFront. Deployed and live.

GOATsolana-MCP — First MCP server connecting Claude with the Solana blockchain · 18 GitHub stars

github.com/FarseenSh/GOATsolana-mcp

- 18 read-only tools across data, predictions, security, and visualization; built on Anthropic's MCP protocol + GOAT SDK in TypeScript. **1st Place ($1,000)** at the Solana Global AI Hackathon.

AWARDS & RECOGNITION

- **The Efficiency Enigma** — Cohere Labs, Expedition Tiny Aya 2026 (Mar 2026). Awarded by Marzieh Fadaee and Tom Hosking. Verified credential available.
- **1st Place, Solana Global AI Hackathon** — $1,000 prize for GOATsolana-MCP.

TECHNICAL SKILLS

Research areas: Mechanistic interpretability, Sparse Autoencoders (JumpReLU, BatchTopK), encoder-side SAE adaptation, representation engineering, LLM evaluation, RAG, RL agents, agent alignment, emergent misalignment.

ML/AI: PyTorch, SAELens, Unsloth, HuggingFace Transformers, LoRA/PEFT, GRPO, contrastive learning, AWS Bedrock (Kimi, DeepSeek, GLM, MiniMax, Cohere Embed).

Languages: Python, TypeScript, JavaScript, Solidity, SQL.

Backend / Web: FastAPI, asyncpg, pgvector, React 19, Vite, Tailwind, Node.js.

Cloud / Infra: AWS (Bedrock, Cognito, RDS, S3, Step Functions, DynamoDB, EventBridge, EC2, CloudFront), AMD MI300X / ROCm.

Protocols: Model Context Protocol (MCP), GOAT SDK.

EDUCATION

Lovely Professional University (LPU), India — Master of Computer Applications (MCA) · CGPA 8.6 / 10 · 2022 – 2024