Malachi Mabie

AI/ML Systems Engineer

Carpinteria, CA
malachi@outlook.com
malachimabie.com
in *mlmabie*

Designing simple affordances in complex environments.

Professional Experience

Apr 2024–Mar 2025 **Senior AI Research Engineer**, *RAM Laboratories, Inc.*, San Diego, CA

- Principal Investigator, T-UEBA ($250k SBIR): led architecture, eval design, cyber ontology/schema development, and constrained deployment strategy for tactical Zero Trust behavior analysis
- Designed CC-GAT with Hopfield conditioning and conformal prediction; delivered 95% accuracy, 0.887 F1, 15.2ms/graph, 15MB model size, and <3% FPR
- COG-SAFE (DARPA FACT): built an LLM harness for cognitive security and accountable agent transactions; Georgia Tech collaboration; represented RAM Labs at Google
- SABRE / pySABRE: tactical edge security work spanning Dockerized virtual simulations, NDN-stack integration, and a solo-built sharded Byzantine fault tolerance simulation/testbed using PBFT, HotStuff, and Streamlet for the C++ EMANE runtime
- Converted research churn into concrete systems decisions around data generation, evaluation, runtime constraints, and deployable model footprints

Apr 2022–Apr 2024 **AI/ML Research Engineer**, *RAM Laboratories, Inc.*, San Diego, CA

- DARPA AIxCC: built an AI vulnerability harness orchestrating RoBERTa, GraphCodeBERT, CodeT5++, and LLaMA QLoRA with custom sampling/decoding and a static-analysis → decompilation → runtime-verification pipeline
- DLR-TV / AIxCC / DL-Patcher: ran low-resource training and post-training work on 4×T4 GPUs using FSDP, custom flash attention, continued pretraining, rejection sampling, and RL-shaped evaluation for vulnerability detection and repair
- DAICON: distributed inference for heterogeneous edge devices using Fused Tile Partitioning, custom PyTorch hooks, work-stealing scheduling, and NDN-based shard routing
- DL-Patcher: patented automated vulnerability repair via pre-trained language models (U.S. 18/375,839); authored 15+ SBIRs including LEAR, HopModKG, Warfighter AI, and VAAST
- Worked across cybersecurity, CV, distributed systems, and NLP programs, often decomposing ambiguous technical goals into tractable tasks for smaller teams

Dec 2024–Present **AI/ML Systems Consultant**, *Mabie Industries, LLC*, Carpinteria, CA

- Through Mabie Industries, expanded a software-architecture consulting engagement into software architecture, implementation, debugging, deployment, infrastructure, and product design while shipping 6 production systems solo (50k+ LOC), including a thermal CV lead-gen pipeline, multi-tenant marketplace, and webhook-driven back office on $35/month infrastructure supporting $811k ARR potential
- Established technical specifications, cloud architecture, and development standards across serverless and Kubernetes-backed infrastructure so the product surface could iterate rapidly without destabilizing the operational core
- Built an OTP-based agent runtime with signal classification, 3-store memory, MCP-based tool orchestration, dynamic subagent spawning, and controlled execution patterns
- Visiting researcher at DeepShard on constrained inference, edge memory, and model distribution; completed technical evaluation work around decentralized RL training

2013–2022 **Earlier Engineering & Teaching**, *Multiple Organizations*, CA

- **Applied Medical** (2013–2018): robotic tooling for medical-device manufacturing; $100k/month savings; intern at 16 → technician → engineer
- **Inspirit AI** (Stanford, 2021–2022): advanced ML instruction and curriculum in RL, transformers, and GANs
- **Endura Technologies / RoadReader / UCSD Design Lab**: deep-learning data mining/cleaning, Java lab software from chip register maps with Verilog / EE teams, CV sensor fusion, and human-AI interaction research for Hyundai robotaxi and Ford mobility

Education

2021 **B.S. Cognitive Science: ML & Neural Computation**, *UC San Diego*
Reinforcement Learning, Neural Signal Processing, Convex Optimization, Systems Programming, HPC

Technical Expertise

Training
FSDP, Flash Attention
Continued Pretraining
LoRA, QLoRA, Distillation
RL: PPO, DPO, GRPO, DQN, MCTS

Inference & Agents
Distributed Inference, ONNX
PyTorch Hooks, Model Sharding
MCP, Memory Systems, Verifiers
Context Engineering, Evals

Stack
Python, C/C++, Elixir/OTP
TypeScript, PyTorch, PyG
Transformers, TensorRT, CUDA
Docker, Terraform, K8s

"Uncanny ability to bring abstract ideas to reality." — Brett Lewis, Senior Engineering Manager, Applied Medical
"Brings a novel and timely perspective to all projects." — Dr. Robert McGraw, VP, CTO, RAM Labs