ANTHONY MAIO

Staff AI Platform Engineer | LLM Infrastructure & Reliability | Agent Systems & Safety Engineering

anthony@making-minds.ai | (914) 325-2482 | Danbury, CT (Remote US)

linkedin.com/in/anthony-maio | github.com/anthony-maio | huggingface.co/anthonym21 | Portfolio: https://making-minds.ai

ResearchGate | Google Scholar | OrcID: 0009-0003-4541-8515

PROFESSIONAL SUMMARY

Staff+ AI Platform and Reliability Engineer with 20 years of production systems experience across regulated fintech, identity/access management, and high-throughput distributed systems. Currently focused on LLM infrastructure, agent orchestration, model lifecycle management, and AI safety engineering. Track record of measurable production impact: 99.99% login uptime, ~20% auth latency improvement, ~$654K/year infrastructure savings, event-sourced systems sustaining ~5K tx/sec at <10ms latency, and Kafka-based compliance pipelines processing ~5TB/day. Combines research-grade prototyping (model training, evaluation harnesses, preference optimization) with production shipping discipline (CI-gated validation, rollout gates, incident response, observability).

CORE SKILLS

AI Platform & LLMOps: LLM evaluation harnesses, agent orchestration, model lifecycle management, inference infrastructure, regression testing, failure-mode analysis, monitoring/observability, safe rollout measurement, prompt/tool risk controls, sandboxed tool execution, AI platform governance, dataset curation, preference optimization (DPO/RLHF)

ML & Training Infrastructure: Python, PyTorch, Hugging Face Transformers/Datasets/TRL, distributed training (DDP), full fine-tuning, LoRA/PEFT, quantization (4-bit/GPTQ/AWQ), model optimization, Mixture-of-Experts architectures

Distributed Systems & Reliability: AWS (EC2, ECS, Lambda, S3, SQS, Kinesis), Kubernetes, Terraform, CI/CD (GitHub Actions, Jenkins), Kafka, CQRS/event sourcing, PostgreSQL, Redis, reliability engineering, incident response, cost optimization, autoscaling

Security & Compliance: AuthN/AuthZ, OAuth 2.0/OIDC, MFA/2FA, API security, audit/compliance pipelines, SOC 2, ISO 27001, PCI DSS, GDPR, KYC/KYB/AML, regulated environment operations

Languages & Tools: Python, C#/.NET, JavaScript/TypeScript, Bash, SQL, Git, Docker

PROFESSIONAL EXPERIENCE

MAKING-MINDS.AI — Remote

Staff AI Systems Engineer — LLMOps, Agent Infrastructure, Safety & Reliability — *Sep 2024 – Present*

Independent AI platform engineering practice delivering LLM infrastructure, agent reliability tooling, and AI safety evaluation systems. Clients include C-level stakeholders at startups and mid-sized industrial organizations (manufacturing/IoT, mineral processing). All artifacts published as open-source with reproducible training/evaluation configurations.

- Designed and built agent execution and verification workflows treating agents as production actors with explicit latency/cost budgets, deterministic replay/verification gates, and CI-gated trust scoring for tool outputs and generated code.
- Implemented sandboxed, headless tool-execution infrastructure with constrained permissions, auditable failure modes, and documented operational tradeoffs for agent orchestration deployments (Agent Hardening Pack).
- Shipped LLM evaluation and monitoring tooling focused on regression detection and failure classification under operational constraints (throughput/latency/cost), with automated CI pipeline integration for model/prompt/tool change validation.
- Trained and released the Eve-2 model family: a 272M-parameter Mixture-of-Experts base model pretrained from scratch on ~10.5B tokens (FineWeb-edu) using PyTorch Distributed Data Parallel, plus instruction-tuned and task-specialist derivatives optimized for CPU/edge inference. Published weights, configurations, and evaluation artifacts to Hugging Face.
- Built Argos-Swarm, an automated red/blue teaming and orchestration framework for multi-model adversarial safety evaluation, including cross-model epistemic divergence measurement for detecting weak-verified failures in LLM outputs.
- Applied preference optimization (DPO/RLHF-style) to fine-tune GLM-4.7 for domain-specific protocol adherence; documented quantization/pruning tradeoffs under controlled adversarial testing (PV-EAT: persona vector + evolutionary adversarial testing framework).
- Delivered AI product engineering reviews and modernization roadmaps to C-level stakeholders, covering inference infrastructure architecture, model lifecycle management, AI platform governance, and cost optimization strategies.

DRAFTKINGS Remote

Staff Software Engineer, Identity & Access Management Platform *Jan 2023 – Aug 2024*

Technical lead and architect for an 8-engineer IAM platform team supporting authentication, authorization, and compliance infrastructure across DraftKings' consumer and enterprise product lines.

- Led and mentored 8-engineer remote IAM platform team; materially improved delivery throughput through tightened ownership norms, structured review workflows, paired programming practices, and pragmatic process standardization.
- Executed zero-downtime 2FA/SMS provider migration: maintained 99.99% login uptime, improved auth latency ~20%, and reduced annual vendor spend by ~$450K; established organization-wide MFA standards across all platforms.
- Reduced AWS infrastructure spend by ~$204K/year through right-sizing, targeted serverless adoption, Kubernetes HPA autoscaling tuning, and resource utilization optimization across IAM platform services.
- Architected and operated Kafka-based compliance data pipelines (Kafka to Redshift), processing ~5TB/day to support regulatory reporting, audit requirements, and KYC/AML compliance workflows.
- Shipped AI-assisted development tooling (code review automation + knowledge retrieval workflows) adopted across 7 engineering teams; reduced review cycle time ~75% and measurably unblocked feature delivery velocity.

Staff Software Engineer, Trading Systems (New Ventures) *Aug 2022 – Dec 2022*

- Delivered 0-to-1 MVP for a greenfield trading platform; drove technical strategy and architecture decisions across two distributed engineering teams during early-stage product development.
- Built event-sourced CQRS core sustaining ~5K transactions/sec at <10ms p99 latency (PostgreSQL + Kafka), establishing the reliability and performance foundation for the trading platform.
- Designed and implemented self-service experimentation infrastructure that reduced feature launch cycle time from days to <1 hour, enabling rapid product iteration and data-driven decision-making.

BROADRIDGE FINANCIAL SOLUTIONS Remote

Staff Product Engineer, Enterprise FX Platform *Jan 2015 – Aug 2022*

- Led staged modernization from monolithic .NET/WCF architecture to AWS microservices for an enterprise FX platform (UBS partnership; ~$2B AUM context), delivering incremental MVPs while onboarding pilot enterprise clients.
- Architected Kafka-based orchestration spanning ~22 interdependent business processes across 7 financial institutions; designed for consistency/partition tolerance tradeoffs with full operational visibility and monitoring.
- Enabled material revenue impact through reliability improvements, incremental platform adoption, and staged delivery methodology that de-risked enterprise client migrations.

TWOFOUR SYSTEMS Remote

Senior Product Engineer, Trading & Risk Platforms *Jul 2007 – Dec 2014*

- Founding engineer building trading and risk platforms for Tier-1 financial institutions in client-facing, on-site delivery roles; delivered real-time risk/margin capabilities as a key product differentiator addressing identified market gaps.
- Designed and shipped production risk calculation engines processing real-time market data feeds, supporting portfolio-level margin and exposure monitoring for institutional trading operations.

EDUCATION

Bachelor of Science, Computer Science, 2003-2007

Binghamton University — Thomas J. Watson College of Engineering and Applied Science, 3.74 GPA

SELECTED PROJECTS

CoDA-GQA-L: Differential Attention Mechanism for LLM reducing KV-Cache VRAM 10-1,000 + 2 Triton Kernels. (info, paper, code)

Self-Directed Knowledge Acquisition in Agentic Large Language Models: (info, paper)

Agent-Nexus: Discord based multi-agent personal project management: (code)

Safety-Lens: AI Safety Visualization Tool; See *how* models think, not just what they say. (code, paper)

Streamlined Intra-Agent Protocol / Slipstream: 60-80% Agent Coordination Token Reduction Protocol + Models (info, code, paper)

Argos Swam: Automated LLM red/blue teaming solution with an evolutionary adversarial pipeline and swarm verification (code)