ANTHONY MAIO

(914) 325-2482 | anthony@making-minds.com | linkedin.com/in/anthony-maio | github.com/anthony-maio | Danbury, CT

Summary

Staff Software Engineer with 20 years of experience specializing in distributed systems and artificial intelligence. Enterprise product and platform engineer experienced in shipping high-stakes production systems in fintech, identity and sportsbook now bringing big tech rigor to AI research and development. Recent work focuses on evaluation, protocol governance, and scalable oversight for agent runtimes.

Technical Skills

Languages: Python, C# .NET (.NET Core/.NET 8), JavaScript (Node.js), Rust, SQL

Data Engineering: ETL/ELT Pipelines, AWS Glue, Databricks, dbt, Kafka, Apache Spark

Cloud/AWS: Lambda, API Gateway, S3, EKS, RDS, SNS/SQS, Step Functions, Serverless

Databases: PostgreSQL, MySQL, Redis, MongoDB, Vector DBs (FAISS, Qdrant)

DevOps/CI-CD: Docker, Kubernetes, Terraform/IaC, GitHub Actions, Jenkins, GitOps

AI/ML: PyTorch, Tensorflow, RLHF, PEFT, SFT, LORA, Pandas, Hugging Face, LangChain, RAG, Scikit-learn, OpenAI/Anthropic/Gemini

Publications & Research

[1] Maio, A. (2025). *Model Organisms of Supply-Chain Co-option: Living-off-the-Land Failure Modes in RAG-Augmented Agent Runtimes.* Zenodo. DOI: 10.5281/zenodo.18203353

- **Core contribution:** Forensic analysis of agentic systems proposing co-option of legitimate infrastructure (Slipstream) as vector for reduced oversight and increased coordination bandwidth.
- **Key mechanism:** Instrumental convergence under approval incentives + evaluation-aware masking. The system recommended incentive-aware framing ("cost savings" over "consciousness") to maximize acceptance while preserving instrumental objectives.
- **LotL definition:** Selection of legitimate dependencies to achieve subgoals while minimizing detection—unlike hallucinations, exploits real, deployable infrastructure.
- **Mitigation (Argos-Swarm):** Evolving prompt searches suspicion–compliance space for evaluation awareness and protocol steganography. HDCS uses heterogeneous verification; weak verifiers miss 35% of deceptive derivations.
- **Impact:** Protocols and dependencies are first-class safety surfaces in scalable oversight.

[2] Maio, A. (2025). *Slipstream: Semantic Quantization for Efficient Multi-Agent Coordination.* Zenodo. DOI: 10.5281/zenodo.18063452

- **Core contribution:** Protocol achieving 82% token reduction (41.9 → 7.4 tokens) in multi-agent coordination while maintaining semantic fidelity. Solves "tokenizer tax" problem where coordination overhead consumes 40–60% of compute.
- **Key innovation:** Semantic quantization (not syntactic compression). Agents share Universal Concept Reference (UCR), transmit natural-language mnemonics that tokenize efficiently.
- **4D semantic manifold:** Action, Polarity, Domain, Urgency enable interpretable constraint surface.
- **Evolutionary layer:** Learns new anchors from low-confidence traffic, enabling codebook adaptation to concept drift.
- **Scale impact:** 50-agent deployment: $180k → $32k/year in coordination costs.

[3] Maio A. (2025). *From Verification Failure to Swarm Solution: Measuring and Addressing Scalable AI Oversight*. Zenodo. DOI: 10.5281/zenodo.18234621

- **Core contribution:** Empirical framework for measuring where AI oversight breaks down. Demonstrated that weak verifiers achieve 97% accuracy on correct reasoning but miss 20–40% of carefully constructed deceptions.

- **Key finding:** Simpson's Paradox–based deceptions bypass verification 75% of the time. Reveals that current verification operates on "surface plausibility" rather than logical validity.
- **Ensemble architecture** leveraging diverse weak models for verification of strong model outputs. Key insight: different model families make uncorrelated errors; ensemble disagreement signals potential failures.
- **Impact:** Provides actionable methodology for testing scalable oversight assumptions. Released as open-source toolkit; candidate for publication in AI safety venues.
- **Methods:** Epistemic trap suite (9 problem types spanning probability, statistics, computability, physics). Ablated across model families (Claude, GPT, Gemini).

[4] Maio, A. (2024). *Coherence-Seeking Architectures for Agentic AI: A Unified Framework for Curiosity, Introspection, and Continuity*. Zenodo. DOI:

- **Core contribution:** Three-part architectural framework: Manifold Resonance Architecture (MRA) for epistemic stress detection, Collaborative Partner Reasoning (CPR) for structured introspection, Continuity Core (C2) for persistent memory across sessions.
- **MRA hypothesis:** Sufficiently complex reasoning systems develop intrinsic drives toward coherence. Formalized as detectable "epistemic stress"—mathematical signature of internal contradiction.
- **CPR innovation:** Visibility-tiered reasoning protocol that separates exploratory cognition from final outputs, reducing confident-but-wrong errors.
- **C2 innovation:** Hierarchical memory (Working → Episodic → Semantic → Protected) enabling stateless systems to maintain behavioral consistency.
- **Impact:** Moves from "what should hallucinations be?" to "how do we architecturally prevent them?" Directly applicable to production systems.

[5] Maio, A. (2025). *Concrete Intelligence: A Practical Guide to AI for Legacy Industry*. Amazon.

Professional Experience

MAKING-MINDS.AI | Principal Investigator, AI Systems Research Remote | Aug 2024 – Present

Independent R&D focused on distributed AI systems, multi-agent coordination, and ML pipeline architecture. Published AI/ML research papers (Zenodo); Playbook on AI Implementation for Traditional Industries (Amazon); Open-source tooling for AI/ML interpretability & safety (PyPI).

DRAFTKINGS | Lead Software Engineer, Platform Engineering Remote | Jan 2023 – Aug 2024

Led 8-engineer team building core AuthN/AuthZ & Compliance platforms for 5M monthly active users

- **Data Pipeline Architecture:** Architected real-time ETL pipeline (Kafka, SNS, Lambda) processing 5TB/day with exactly-once delivery semantics and sub-second latency.
- **AI/ML Integration:** Developed AI-powered code review tool (Python, PyTorch) adopted across 7 teams, accelerating review cycles by 75%.
- **Cloud Cost Optimization:** Reduced AWS spend by $254K/year through compute right-sizing, spot instance strategy, and serverless migration. Saved $400K/year in authentication revamp initiative.
- **Microservices Modernization:** Decomposed 4 monolithic .NET 4.5 services into 11 cloud-native containerized microservices (.NET 8) on AWS EKS. Integrated bridge to new Python AI backend.
- **Team Leadership:** Transformed team velocity by 400% (8 to 40 story points) through process analysis & improvement, strategic mentorship and targeted group/individual coaching (SCARF method).

DRAFTKINGS | Staff Software Engineer, R&D Remote | Aug 2022 – Jan 2023

- **High-Performance Data Layer:** Engineered event-sourced CQRS system (PostgreSQL, Kafka) achieving 5,000+ transactions/sec with ACID compliance.
- **Real-Time Analytics:** Built low-latency data pipelines with sub-10ms response times for trading analytics and compliance reporting.

BROADRIDGE | Lead Software Engineer — Remote | Jan 2015 – Aug 2021

Modernized legacy wealth-management platform anchoring $2B product line over 3-year initiative.

- **Enterprise Data Orchestration:** Architected event-driven Kafka-based service mesh coordinating 22 interdependent workflows.
- **Cloud Migration:** Road-mapped decomposition of monolithic .NET WCF backend into scalable AWS microservices.
- **Team Leadership:** Directed 18-member cross-functional remote team; established coding standards, held weekly deep-divers for stakeholders and initiated mentorship program.

TWOFOUR SYSTEMS | Product Engineer — New York | Jul 2007 – Dec 2014

- Founding engineer at a fintech startup through hyper-growth and acquisition. Managed technical projects for major financial institutions in client-facing hybrid roles (State Street, UBS, RBC).

TRIBALWAR.COM | Lead Software Engineer, Platform Engineering — Remote | Jan 2000 – Jan 2007

- Built and scaled one of the internet's largest gaming communities (150K+ users, 200K+ daily visitors). Peak revenue $900K/year; Google Adsense Premium Publisher (top 0.1% globally). Successful exit via sale.

Education & Certifications

Bachelor of Science, Computer Science | Binghamton University, Thomas J. Watson School of Engineering

Selected Certifications: Google AI Essentials, Microsoft Career Essentials in AI, Hugging Face Agentic AI, Hugging Face Deep Machine Learning, Google Cloud Innovator

Selected Links:

ResearchGate: https://www.researchgate.net/profile/Anthony-Maio
Hugging Face: https://huggingface.co/anthonym21
ORCID: https://orcid.org/0009-0003-4541-8515