

Daniel Yarmoluk

Agentic AI Architect | Autonomous Multi-Agent Systems | LLM Orchestration and Enterprise Integration

Minneapolis, MN | daniel.yarmoluk@gmail.com

linkedin.com/in/danyarmoluk | github.com/Yarmoluk | pypi.org/user/danyarm | huggingface.co/danyarm | graphifymd.com

PROFESSIONAL SUMMARY

Agentic AI architect who designs and ships autonomous multi-agent systems in production. Architected and deployed agent orchestration where an orchestrating agent queries a domain knowledge graph and dispatches specialized worker agents, each pre-loaded with a task-scoped subgraph so the agent reasons over structured context instead of retrieving after the fact. Built 6 production Model Context Protocol (MCP) services and 12 published Python packages delivering that context to any LLM orchestrator — LangChain, LangGraph, AutoGen, CrewAI, Claude, OpenAI. Deep hands-on Generative AI and Retrieval-Augmented Generation (RAG) engineering: built and benchmarked a full RAG pipeline (FAISS and Chroma vector stores, sentence-transformers embeddings, chunking and top-k retrieval) against a typed knowledge graph across 44 domains and 7,758 queries. 15+ years enterprise architecture delivering AI for UnitedHealthcare, Best Buy, Cargill and Salesforce CRM platforms across AWS, Microsoft Azure and Google Cloud. Patent pending. Adjunct professor of graduate AI systems.

PROFESSIONAL EXPERIENCE

Graphify.md — Minneapolis, MN

Founder and Principal Engineer, AI Infrastructure | 04/2026 – Present

- Design and build the agent harness: tool registration and discovery, output schema validation, task-scoped context assembly, orchestration and worker dispatch, guardrails and entitlement gating, retry and fallback, and harness-level observability.
- Design agentic architectures for autonomous task execution: multi-agent orchestration, agent-to-agent coordination, tool/function calling, reasoning and planning loops, and API-driven integration with enterprise systems including Salesforce CRM.
- Integrated production MCP services with Microsoft Semantic Kernel via the MCPStreamableHttpPlugin connector — all 9 tools register as Semantic Kernel functions with no framework-specific server code. The same servers are consumed unchanged by LangChain, LangGraph, AutoGen, CrewAI and Claude, because the integration happens at the protocol layer rather than per framework.
- Authored and published A2A (Agent-to-Agent) agent cards for 3 production services, advertising skills, authentication, x402 payment terms and a machine-readable economics block — token cost, measured answer quality and a break-even input price — so a calling agent can compute expected value before invoking rather than trusting a claim.
- Own agent distribution: published 12 Python packages across 100+ releases and listed the servers in public MCP registries. Instrumented the handshake to measure consumption and identified 27 distinct directory, registry and reliability crawlers indexing the services daily — turning distribution from an assumption into a measured channel.
- Built machine-payable agent infrastructure: x402 / HTTP 402 payment-required flow with EIP-3009 signed authorizations, Coinbase Developer Platform facilitator authentication, Base network settlement, and ERC-8004 agent identity — and documented honestly where no mainnet facilitator existed for the required scheme.
- Own agent performance, latency, cost efficiency and governance — per-call cost modelling, token-efficiency measurement, rate limiting and quota enforcement, structured telemetry, and production monitoring of deployed agent services.
- Designed, built, deployed and operated 6 production Model Context Protocol (MCP) servers serving typed knowledge graphs to AI agents, and published 12 Python packages to PyPI across 100+ releases — owning the complete lifecycle from protocol design through incident response. All of it built solo, first package published within weeks of starting.
- Implemented the full MCP protocol surface in Python: JSON-RPC 2.0 initialize handshake with clientInfo and protocolVersion negotiation, tools/list, tools/call, resources/list, prompts/list, and notifications, with tool input and output schema validation.
- Diagnosed and fixed a production transport defect where every external MCP client received HTTP 421 Misdirected Request: DNS-rebinding protection remained active with an empty host allow-list behind a reverse proxy, rejecting all traffic before it reached a tool. Implemented explicit transport security settings with configurable allowed hosts.
- Built the metering and monetization layer for MCP: classified protocol methods as metered versus unmetered so the three-request handshake no longer consumed a client's entire free tier, implemented time-windowed free access with a call ceiling, per-IP and per-/24-subnet rate limiting, license key validation, and HTTP 402 Payment Required responses with signed payment challenges.
- Instrumented MCP servers with product analytics using PostHog's MCP event schema — capturing tool calls, client identity from the initialize handshake, latency, errors and funnel conversion — and corrected three instrumentation defects that had made usage unattributable to any caller.
- Implemented Agent-to-Agent (A2A) and machine-payment integrations: x402 / HTTP 402 request flow with EIP-3009 signed authorizations, Coinbase Developer Platform facilitator authentication, Base network settlement, and ERC-8004 agent identity — and documented the honest limitation that no mainnet facilitator existed for the required scheme.

- Built a conjunctive query operator for knowledge graphs exposed as a single MCP tool — returning concepts reachable from every anchor as an exact set intersection over declared edges, with typed relation paths, AND/OR modes, and a per-step cardinality trace that makes retrieval auditable.
- Designed a SHA-256 per-node source provenance system so any answer can be traced to the exact bytes of its source, including canonical-text hashing for CDN-rendered documents whose raw bytes carry rotating build tokens, and a verification tool that reports honestly when a source cannot be independently verified.
- Benchmarked Compressed Knowledge Graph retrieval against RAG and Microsoft GraphRAG across 44 domains and 7,758 queries: macro-F1 0.471 versus 0.123 (RAG) and 0.120 (GraphRAG), a 3.8x improvement; 269 versus 2,982 tokens per query, an 11x reduction; multi-hop F1 rising to 0.772 at 5 hops while RAG remained flat near 0.17. Published the dataset and harness openly.
- Built AI evaluation infrastructure: a reproducible retrieval-evaluation harness plus an independent evaluation against FrontierFinance, Samaya Research's open finance benchmark (220 queries, 11,543 rubric items, LLM judge panel) — evaluating against a third-party rubric rather than only a self-authored one.
- Authored and maintained 300+ domain knowledge graphs with typed relations and per-node source provenance across healthcare, financial regulation, data engineering, and AI infrastructure.
- Patent pending — three provisional applications filed 2026 covering knowledge graph architecture, encoding pipeline, and context infrastructure.

Slalom Consulting — Minneapolis, MN

Principal — Solution Owner, AI and Digital Transformation | 09/2023 – 04/2026

- Architected enterprise AI and digital transformation programs for Fortune 500 clients including UnitedHealthcare (clinical AI pipelines), Best Buy (retail AI), and Cargill (supply chain AI).
- Led cross-functional AI delivery across engineering, data science and business stakeholders, translating strategy into production system architecture and deployment roadmaps.
- Designed AI workflow orchestration and agent coordination patterns for enterprise deployment, including production readiness, monitoring and evaluation frameworks for regulated environments.
- Managed technical vendor relationships and platform integration across Anthropic, OpenAI, Azure OpenAI Service on Microsoft Azure AI, and AWS SageMaker — owning technical due diligence and build-versus-buy recommendations assessed on latency, cost, compliance fit and roadmap.
- Operated inside complex, matrixed professional services organizations serving financial services, healthcare and regulated clients — able to influence architecture decisions across business lines without direct authority, and to translate strategic priorities into actionable architecture decisions with measurable business impact. Recognized internally as a technical authority and thought leader on enterprise AI.

West Monroe Partners — Minneapolis, MN

Lead Solution Architect, Industrial IoT and Commercial AI | 2021 – 2023

- Led AI and Internet of Things (IoT) architecture for industrial and commercial clients: production machine learning pipelines, real-time data systems, and multi-system orchestration.
- Designed production deployment architecture for AI systems in operational environments, covering reliability, monitoring and performance under load.

ATEK Access Technologies — Minneapolis, MN

IoT and Data Science Practice Leader | 2016 – 2020

- Built the data science and IoT practice from zero, including AWS, Siemens and IBM strategic partnerships and production machine learning deployment in industrial field operations.
- Led sensor data orchestration across distributed field devices, including real-time data pipeline design and system integration architecture.

University of St. Thomas — Minneapolis, MN

Adjunct Professor, Graduate AI Systems (SEIS 666) | 2018 – Present

- Eight years accelerating large language model (LLM) and knowledge graph capability for senior product and engineering leaders. Students are practicing technology executives and staff-level engineers who apply the material directly to live production initiatives at their own companies.
- Designed and continually rewrite graduate curriculum spanning production LLM systems, agent orchestration design, knowledge graph infrastructure, enterprise AI evaluation, and digital transformation — updated each term as the capability frontier moves.
- Enterprise AI adoption and enablement at leadership level: technical upskilling of senior technologists, change management, and closing the gap between what a system can do and what an organization will actually put into production.

BENCHMARK — PUBLISHED AND REPRODUCIBLE

Compressed Knowledge Graph versus Retrieval-Augmented Generation (RAG) and Microsoft GraphRAG. Version 0.6.2, locked: 44 domains, 7,758 queries. Dataset and evaluation harness published openly (CC-BY-4.0) so results can be independently re-run.

- Answer quality: macro-F1 0.471 (CKG) versus 0.123 (RAG) and 0.120 (GraphRAG) — 3.8x.
- Context efficiency: 269 versus 2,982 tokens per query — 11x fewer tokens.
- Multi-hop reasoning: F1 rises to 0.772 at 5-hop depth while RAG stays flat near 0.170.
- Auditability: every answer traces to a declared relationship and, where source provenance is present, to a SHA-256 hash of the source bytes.
- Reproducibility: github.com/Yarmoluk/ckg-benchmark and huggingface.co/danyarm — clone and re-run.

PUBLICATIONS AND RECOGNITION

- Publication in preparation: a systematic benchmark of compressed knowledge graphs for enterprise AI, in collaboration with Dan McCreary, former Head of AI at TigerGraph.
- Media: featured in CIO Dive on enterprise context limitations; guest on the AI and Data Horizons podcast; host of All Things Data (1M+ downloads). Benchmark cited independently at Data Tech 2026.

EDUCATION

- Graduate Degree, Data Science — University of St. Thomas
- Master of Business Administration — Loyola University Chicago
- Bachelor of Arts, International Affairs — University of Texas