

Daniel Yarmoluk

Forward Deployed Engineer | Enterprise AI Deployment | 0 to 1 Under Ambiguity

Minneapolis, MN | daniel.yarmoluk@gmail.com

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

PROFESSIONAL SUMMARY

Forward deployed engineer who embeds with customers and ships production software against ambiguous requirements. Built a company 0 to 1 in twelve months — 12 Python packages published across 100+ releases, 6 production services, 307 domain knowledge graphs and an open, re-runnable benchmark — making deliberate standardization-versus-customization decisions rather than defaulting to bespoke work: a reusable typed-graph substrate replaced per-customer retrieval builds. Delivered production AI inside Fortune 500 environments — UnitedHealthcare, Best Buy, Cargill — owning outcomes from first workshop through deployed system and writing the integration code rather than handing off a diagram. Codifies successful patterns into reusable primitives: packages, protocol servers, and 84 authored agent skills, so each engagement starts further along than the last. Writes and reviews production-grade Python; ships, meters, secures and monitors live services. Debugs production incidents in unfamiliar systems under time pressure. Eight years as a graduate professor teaching working technologists to deploy and adopt AI inside their own organizations. Publishes field findings honestly, including a public correction to his own benchmark. Patent pending.

PROFESSIONAL EXPERIENCE

Graphify.md — Minneapolis, MN

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

- Act as technical lead on ambiguous, high-stakes deployments, owning problems end-to-end from definition to resolution and owning both the technical and the business outcome.
- 0 to 1 under ambiguity: built the entire product surface from nothing — retrieval architecture, 6 production Model Context Protocol (MCP) services, 12 published packages, provenance and payment layers, and the benchmark measuring whether any of it worked.
- Made the standardization-versus-customization call explicitly: chose a portable typed-graph substrate reusable across 307 domains over bespoke per-customer retrieval, then measured the trade-off openly against RAG and Microsoft GraphRAG across 44 domains and 7,758 queries.
- Built and operate the agent harness itself — the runtime between a model and what it acts on: tool registration and discovery, schema validation at the boundary, task-scoped context assembly, worker dispatch, entitlement gating, retry and failure isolation, and the telemetry to see what actually happened.
- Codified patterns into deployment primitives that create leverage: installable packages, protocol servers consumed unchanged by every major agent framework, reference integrations, and 84 authored, versioned agent skills.
- Write and ship production code on both sides of the stack: Python services plus JavaScript serverless functions — including a JSON-RPC MCP server implemented in JavaScript and a Stripe webhook with signature verification — and the interfaces on top, with interactive graph visualizations hand-built in canvas and SVG rather than pulled from a charting library.
- Debug production incidents in unfamiliar systems under time pressure: diagnosed a transport defect rejecting 100% of external clients while health checks stayed green, a tool-schema mismatch silently breaking every caller behind HTTP 200s, and an authorization bypass caught before it shipped — each found by reading the protocol and the raw data, not the dashboard.
- Player-coach by practice: Principal leading cross-functional Fortune 500 delivery, and eight years teaching mid-career engineers and architects to deploy and adopt AI in their own organizations — the enablement half of forward deployment, not just the build.
- Translate field findings into product and research feedback, including publishing a reconciliation correcting his own cost figures when the original comparison proved flawed. Numbers that cannot be defended are worse than no numbers.
- 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