Yuriy S.
Architect AI Solution Architect / Engineering Manager

SUMMARY

- AI Solution Architect and Engineering Manager with 20+ years of experience in software engineering and delivery.
- Owns production RAG architecture on AWS using Python, LangChain, Weaviate, OpenAI, and AWS Bedrock, covering retrieval strategy, embeddings, privacy boundaries, quality calibration, and LLM cost monitoring.
- Designed multi-agent workflows with autonomy boundaries, governance controls, CI policy gates, and test-first enforcement.
- Led cross-functional teams and delivery from proof of concept to production, including stakeholder alignment and architecture decisions.
- Experienced in microservices, enterprise integrations, security compliance, and cloud-native systems.

TECHNICAL SKILLS

Main Technical Skills	Python, LLM, LangChain
AI Tools & Assistants	Devflow
Programming Languages	Java, Python
AI & Machine Learning	AI Governance, CrewAI, GPT-4o, LangChain, LLM, Microsoft Presidio, MVTec AD, OpenAI, OpenCV, RAG, RT-DETR, Semantic Search, YOLO
Java Libraries and Tools	Apache AXIS, Apache Camel, EJB, Google Web Toolkit, ILOG JRules, Java Persistence API, Java SE, Java Server Pages (JSP), JAXB, JDBC, JMS, JPA, JSON
Mobile Frameworks and Libraries	Flutter
JavaScript Frameworks	Google Web Toolkit, GWT, React
Java Frameworks	GWT, Hibernate, Spring, Spring Boot
Security	AI Governance, Microsoft Presidio, OWASP, SOC 2
Databases & Management Systems / ORM	Couchbase, ELK stack (Elasticsearch, Logstash, Kibana), Hibernate, Microsoft SQL Server, Oracle 10g, Oracle Database, SQL, Vector DB
UI Frameworks, Libraries, and Browsers	HTML, MDB
Cloud Platforms, Services & Computing	AWS
Amazon Web Services	AWS Bedrock, AWS Lambda, AWS S3

Methodologies, Paradigms and Patterns	Agile, Architecture Decision Records, Devflow, GDPR, Kanban, microservices, REST, SAFe, Scrum, SDLC, SOAP, TDD, UML
Platforms	Amdocs CRM
Web/App Servers, Middleware	Apache Tomcat, IBM WebSphere Application Server
QA, Test Automation, Security	Automated testing, Code Review, Penetration testing
Deployment, CI/CD & Administration	CI/CD, Devflow, Jenkins, New Relic
Virtualization, Containers and Orchestration	Docker, Kubernetes, Terraform
Version Control	Github Actions
Operating Systems	Oracle Solaris
Logging and Monitoring	Prometheus
Message/Queue/Task Brokers	RabbitMQ
File Systems, Storage	XML
Other Technical Skills	Amazon API Gateway, C4 Model, Claude Code, Data pipeline design, JBoss Application Server, jBPM, Jolt, Observability, Oracle WebLogic Server, SDD, Weaviate

WORK EXPERIENCE

NDA, Engineering Manager

Jan 2026 - Present

Forge Sentinel

Forge Sentinel is an AI-powered manufacturing quality-control system that detects surface and component defects in real time using a fine-tuned YOLOv8 model trained on MVTec AD industrial benchmark data. Detection events are persisted in a knowledge store and queried through a RAG layer, enabling non-technical operators to query inspection history in natural language.

Delivery & Team Leadership

- Led a cross-functional working group of 5 engineers across computer vision, machine learning, backend, and data through investigation and implementation, from problem framing to a working system.
- Led technical direction and delivery, defined the AI system architecture, and set formal quality acceptance gates of 81% precision, 85% recall, and 0.82 mAP@0.5.
- Coordinated engineering decisions on frame processing, severity classification, and the detection-event schema, including production-line zone awareness.
- Governed the RAG-based inspection-query capability for non-technical operators.

Solution Architecture

- Ran the detection technology evaluation, benchmarking YOLO-family models against RT-DETR on MVTec AD. Selected fine-tuned YOLOv8 and recorded the trade-off and rejected alternatives as architecture decision records.
- Architected the RAG query layer over a dual-collection Weaviate knowledge store, routing between structured filters, semantic search, and LLM trend reasoning.
- Designed an offline-first edge detection layer with no network dependency on the critical path and a graceful degradation ladder for the cloud intelligence layer.
- Defined and validated measurable non-functional targets of 100 ms inference latency per frame and a 93% defect-catch rate on the validation set.

AI-Augmented Delivery

A delivery process analysis and improvement program for a product team, built on Devflow, the candidate's own AI development framework with an open-source core. The goal was to capture AI's speed without losing engineering control, making delivery faster and more predictable while keeping code quality and accountability verifiable.

Delivery & Team Leadership

- Applied combined Delivery Manager, Technical Lead, and Scrum Master experience across a 4–6-person product team to identify and resolve delivery risks and process gaps introduced by AI adoption.
- Established AI governance as a delivery-control mechanism and enforced test-first automated testing, human code reviews, and AI-assisted security scanning as shift-left delivery quality gates.
- Adapted Scrum to AI-accelerated development through refinement-readiness analysis, ceremony support, and honest team-level delivery forecasting, restoring predictable, forecastable delivery.
- Reported 37% faster delivery, a 35% cycle-time reduction, and a 25% drop in review-stage defects. AI usage changed from individual improvisation into a governed, repeatable, auditable team delivery workflow.

Solution Architecture

- Architected the governance layer as a deterministic control plane outside the AI reasoning loop. Defined trust-zone policies specifying where AI may generate code and where human authorship is enforced, with gates at CI merge.
- Designed multi-agent orchestration with role-scoped agents, explicit autonomy boundaries, and comprehension gates before code review, treating controllable autonomy as an architectural property rather than a prompt instruction.
- Established architectural traceability through architecture decision records and a structural change ledger, embedding test-first enforcement and AI-assisted security scanning in the framework runtime.

AI-Augmented Delivery: Devflow, Claude Code, multi-agent workflows, SDD, TDD, GitHub Actions. Forge Sentinel: Python, YOLO-family models, YOLOv8, RT-DETR, OpenCV, MVTec AD, Weaviate, LangChain, GPT-4o, Docker. Additional employment tags: AI Governance, AI Integration.

Forte Group, Project Director/Delivery Manager

Apr 2024 - Jan 2026

CVision

CVision is a production LLM-powered search platform replacing keyword-based HR tools with a natural-language interface over people-management resources. It implements a multi-agent RAG pipeline on a Weaviate vector store with high-dimensional OpenAI embeddings, deployed as a microservice ensemble on AWS with structural privacy compliance.

Delivery & Team Leadership

- Owned end-to-end delivery from proof of concept to production-grade microservice architecture, managing the technical transition while maintaining delivery continuity and stakeholder confidence.
- Directed a cross-functional engineering group across backend, machine learning, and data, and owned stakeholder alignment on scope, cost, and compliance trade-offs.

Solution Architecture

- Owned retrieval strategy, data-pipeline design, privacy architecture, and model selection throughout delivery. Adopted a multi-algorithm RAG approach to balance precision and recall while weighing cost, compliance, and business requirements.
- Designed the document-intelligence pipeline with semantic chunking, OpenAI embeddings, Weaviate as the production vector store, LLM-based reranking, and final relevance scoring.
- Architected privacy as a structural boundary. GDPR-compliant PII anonymization removes sensitive data at the pipeline boundary before any external LLM call, with enforced audit trails.
- Built production observability into the design through continuous quality calibration of retrieval precision, answer relevance, and faithfulness, plus LLM call tracing, inference latency, and cost monitoring, providing operational and budget visibility.

Tidal

Tidal is Redwood Software's mission-critical workload automation platform for centralized orchestration of hybrid business processes. Directed an enterprise automation program involving 50+ engineers.

- Directed end-to-end delivery aligned with Redwood Software standards and supported teams in architectural reviews covering design decisions, scalability, and trade-offs.
- Managed and coordinated Development, QA, DevOps, Security, Product Support, and Product Owner teams in a unified Scrum environment.
- Defined and tracked KPIs for engineering productivity, delivery velocity, release success, and product quality. Owned budget and resourcing decisions and managed risks proactively across releases.
- Owned SOC 2 audit compliance, coordinating regular SAST and DAST security scans, penetration testing, periodic audits, and external audits.
- Served as the delivery point of contact for account management and client stakeholders, managing expectations, escalations, and status reporting.
- Facilitated agile ceremonies for continuous improvement.
- Managed performance reviews, salary adjustments, one-to-one mentoring, team development planning, and the hiring lifecycle.

CVision: Python, OpenAI, high-dimensional OpenAI embeddings, Weaviate, LLM, RAG, vector databases, ML pipeline integration, LangChain, CrewAI, Prompt Engineering, Microsoft Presidio, AWS Bedrock, AWS, Docker, microservices. Tidal: Scrum, SOC 2, SAST, DAST, penetration testing. Additional employment tag: Java.

TalkRemit, Engineering Manager/Office Director

Jun 2020 - Apr 2024

TalkRemit

A digital banking and financial-services platform comprising an AWS-based ecosystem with web, native, and Flutter mobile applications, supporting eWallet, multicurrency cards, and flexible pay-in and pay-out options.

- Scaled the fintech platform to 200K+ users with 99.9% uptime.
- Established and optimized delivery processes, introducing Scrum practices and structured release management across multiple teams.
- Oversaw end-to-end SDLC and delivery governance, with clear environment flow and quality gates for production readiness.
- Directed product support operations and incident management, including knowledge-base development, out-of-hours response, and SLA adherence for a live financial platform.
- Served concurrently as Director of the Ukraine office, overseeing local operations, budget, people management, and organizational alignment with company-wide goals.
- Led OKR implementation and review cycles, team risk management, and one-to-one mentoring.
- Participated in system design discussions, architectural evaluations, and integration decisions to balance delivery speed with long-term maintainability.
- Partnered with external recruiting services, conducted technical interviews, and managed the full hiring lifecycle.

AWS, web applications, native mobile applications, Flutter, Scrum.

Intellias, Delivery Manager

Jul 2019 - Apr 2020

AmDocs

A B2B enterprise telecom solution involving 12 teams using SAFe, built on service-oriented architecture and cloud-native technologies. The CV describes the role as Delivery Manager and Technical Scrum Master.

- Set up and managed Scrum processes in two development teams, aligning delivery practices with SAFe.
- Directed planning, development, and delivery processes in close collaboration with the architecture team.
- Contributed to architectural and technical decision-making, ensuring delivery feasibility while maintaining scalability and compliance.
- Managed risks proactively and represented teams in Scrum-of-Scrums and SAFe synchronization meetings.
- Fostered professional knowledge sharing across teams through guilds and collaborative forums.

SAFe, Scrum, SOA, cloud-native technologies, microservices, Docker, Kubernetes, Jenkins CI/CD, Java 8, React, Couchbase, Prometheus.

Technical Lead, Scrum Master - Ciklum, Technical Manager

May 2015 - Jul 2019

Thomas Cook Integration

An enterprise integration platform based on microservices and enterprise integration patterns for global operations.

- Directed integration project management, including timelines, estimates, SLA negotiations, and risk mitigation.
- Provided technical leadership for Java development teams, including solution design, architectural reviews, deployment improvements with DevOps, and production support.
- Actively contributed to the Integration Competency Center and open-source projects, including msb-java.
- Facilitated Scrum adoption, backlog prioritization with Product Owners, Scrum-of-Scrums, and inter-team collaboration.
- Conducted one-to-one meetings and guided team members through personal development plans and technical mentorship.

Microservices, enterprise integration patterns, RabbitMQ, Java 8, Spring Boot, Apache Camel, Jolt JSON, REST, SOAP, ELK Stack, New Relic, Rackspace Cloud, Terraform, Scrum.

Daxx, Java Team Lead

Jun 2010 - Feb 2015

GT-WOZ

A multi-tiered web application for real-estate registration and the entire property-tax process. Used web services and an enterprise service bus to integrate with national and other external services. The internal business workflow model was implemented with the jBPM workflow engine.

- Managed the team's development process.
- Handled technical communication with customers.
- Provided technical leadership.
- Performed Java development.

Java, EJB 3.0, JPA, Web Services, REST, SOAP, ESB, Hibernate, GWT, jBPM, JBoss Enterprise SOA Platform 4.3, Oracle 10g.

GlobalLogic, Team Lead

Mar 2006 - Apr 2010

NetAnalyst

NetAnalyst is a test and management Operations Support System used for Quality of Service testing and monitoring of IP packet and TDM circuit networks supporting VoIP, IPTV, digital video, DDS, SONET, and SDH/PDH. Served as Technical Lead / Team Lead.

- Organized development and testing processes.
- Handled technical communication with customers.
- Developed technical rules and framework classes.

Communication Engine

An open, flexible software platform enabling wireless and wireline operators to create, package, deploy, and manage interactive IPTV, time-shifted broadcast TV, video-on-demand,

games, and third-party premium content services. Served as Technical Lead / Java Senior Developer.

- Designed and implemented an asynchronous interaction model using JMS and web services.
- Worked on database communication and the data-object layer.

NetAnalyst: Java SE, ILOG JRules, TL1 command language, WebLogic, Solaris.
Communication Engine: EJB 2.1, SLSB, MDB, JMS, Web Services, Axis, JAXB, JDBC, IBM WebSphere, Apache Tomcat, Oracle 10g, UML.

EDUCATION

- National Technical University of Ukraine "Kyiv Polytechnic Institute"
  Master's degree in Computer Science
- Institute of Investment Management
  Specialist in Enterprise Management
- Kyiv Business School
  Specialist; Enterprise Management

CERTIFICATION

- **Model Context Protocol: Advanced Topics**
  LinkedIn source details:; Original qualification period: Jan 2026 – Present
  Jan 2026
- **Retrieval Augmented Generation (RAG)**
  LinkedIn source details:; Original qualification period: Sep 2025 – Present
  Sep 2025
- **Generative AI with Large Language Models**
  LinkedIn source details:; Original qualification period: Jul 2025 – Present
  Jul 2025
- **Building Systems with the ChatGPT API**
- **Practical Application of Gen AI for Project Managers**
  LinkedIn source details:; Original qualification period: Jul 2025 – Present
  Jul 2025
- **Certified ScrumMaster® (CSM®)**
  Expired in March 2021.
  Mar 2019 – Mar 2021
- **Certified Scrum Product Owner® (CSPO®)**
  Expired in March 2021.
  Mar 2019 – Mar 2021
- **Developing Secure Software**
  LinkedIn source details:; Original qualification period: Oct 2024 – Present
  Oct 2024
- **Learning the OWASP Top 10**
  LinkedIn source details:; Original qualification period: Oct 2024 – Present
  Oct 2024
- **Talking to AI: Prompt Engineering for Project Managers**
  LinkedIn source details:; Original qualification period: Sep 2025 – Present
  Sep 2025
- **Generative AI Applications with Amazon Bedrock**
  LinkedIn source details:; Original qualification period: Nov 2025 – Present
  Nov 2025

- **AI Enhancement with Knowledge Graphs - Mastering RAG Systems**
  LinkedIn source details:; Original qualification period: Nov 2025 – Present
  Nov 2025