# JEREMY VELEBER

Principal Engineer • AI-Accelerated Systems Design • Modernization Architect

Lynnwood, WA • 425-778-9588 • resume@jeremyveleber.com • www.JeremyVeleber.com

## PROFESSIONAL SUMMARY

Principal Engineer with **25+ years** of experience designing and modernizing large-scale systems across enterprise, retail, telecom, and cloud environments. Specializes in ground-up architecture, AI-accelerated engineering, and high-velocity modernization that removes months of friction from engineering organizations. Proven track record of building self-healing platforms, zero-touch deployment pipelines, and Standards-as-Code frameworks that scale across teams.

## CORE COMPETENCIES

**AI & ML Engineering:** RAG systems • ChromaDB • sentence-transformers • Maximal Marginal Relevance optimization • Prompt engineering (few-shot, chain-of-thought, structured outputs) • Embedding model selection & benchmarking

**Backend & Data:** Python • Java • C# • Spring Boot • Kafka • SQL/NoSQL • PostgreSQL • RESTful API design

**Cloud & Infrastructure:** AWS (ECS, EKS, Fargate, Lambda, API Gateway, RDS, DynamoDB, ElastiCache, S3, CloudFormation, CloudWatch) • GCP (Cloud Run, Cloud Build) • Azure • Docker • Kubernetes • Istio • Helm

**DevOps & Velocity:** CI/CD (GitLab, Jenkins) • Self-healing infrastructure • Observability • Zero-downtime deployments

**Architecture:** Microservices • Event-driven systems • API modernization • Data contract design

## AI ENGINEERING & APPLIED ML

### Production RAG System

**claude-code-search** — Semantic code retrieval system for AI-assisted development:

- Built production RAG system using **ChromaDB vector storage** with sentence-transformers embeddings
- Benchmarked **12+ HuggingFace models** for code similarity optimization across multiple dimensions (semantic accuracy, inference speed, memory footprint)
- Optimized retrieval quality using **Maximal Marginal Relevance (MMR)** to balance semantic relevance with result diversity—prevents Claude from reading redundant/irrelevant code in RAG context
- Implemented document chunking, ingestion pipeline with open source release and active developer user base

- Evaluating **LangChain for agent orchestration** and **RuVector for self-learning indexes** in future iterations

### AI-Accelerated Development Workflows

**Advanced Prompt Engineering** for system design, architecture analysis, and code generation:

- **Few-shot prompting:** 3-5 diverse examples for pattern teaching, achieving 20-60% quality improvement on complex tasks
- **Chain-of-thought reasoning:** Step-by-step decomposition for complex technical decisions
- **Structured outputs:** XML-tagged instructions for Claude, JSON schemas for GPT models
- Applied role definition, context ordering, and constraint-based solution narrowing techniques

**Production Deployment at Starbucks:**

- Achieved **high-accuracy C# → Java Spring Boot translation** using Microsoft Copilot, eliminating manual translation risk
- Built AI-driven story generation to analyze legacy APIs and guarantee **100% data parity** during migration
- Maintained backward compatibility across all migrated APIs

### Production Full-Stack Systems

**WA Creel** — Cloud-native web application:

- Deployed on **Google Cloud Run** with Docker containerization and automated CI/CD pipeline (Cloud Build)
- Designed RESTful API, data processing pipeline (13 years survey data), interactive visualization (Leaflet, Chart.js)
- Refactored **2464-line monolith into modular architecture** (300-line modules, single-responsibility design)
- Live production system: **wa-creel.jeremyveleber.com**

## PROFESSIONAL EXPERIENCE

**Sirrus7** — **Principal Engineer / Modernization Architect** *Feb 2025 – Present*

*Engineering Velocity & Infrastructure Modernization (T-Mobile, Starbucks, Educational Game Platform)*

- Reduced deployment readiness from **6 months to under 1 week** by designing standardized Kubernetes + Istio platform with automated health checks, security baselines, and GitLab-driven release automation
- Cut developer onboarding from **2-3 weeks to 1-2 days** by building unified, self-documenting environment with baked-in standards and tooling
- Eliminated **10-person release calls** by implementing zero-touch, self-healing deployment pipelines
- Recognized by T-Mobile leadership as the **"ideal reference profile"** for Autofailover engineering engagement
- Designed North Star data model for Starbucks Menu Manager (Promotions and AEM Image data), establishing long-term architecture
- Praised by Starbucks engineering leadership for taking clear ownership of system/data design and driving cross-team alignment

**Ascendion (Starbucks)** — **Senior Manager / Lead Software Design Engineer** *Aug 2024 – Feb 2025*

- Designed unified data integration platform consolidating product data from multiple legacy systems into modern interfaces
- Shifted Starbucks from data-copying patterns to **single-source-of-truth architecture**, improving consistency and reliability
- Led modernization of transaction processing systems with **zero-downtime migration** strategies
- Mentored senior engineers in MongoDB, Kafka, and modern design principles

**Nortal** — **Lead Software Design Engineer** *Nov 2018 – Apr 2024*

- Modernized enterprise systems into hybrid legacy/modern operational models
- Implemented Kubernetes orchestration and microservices to increase throughput and reliability
- Integrated advanced auth systems (OKTA, OAuth, JWT, SAML)
- Led migration of **hundreds of APIs** to modern frameworks with zero downtime

**F5 Networks** — **Senior Java Developer** *2015 – 2018*

- Redesigned state machine software for improved responsiveness and scalability
- Built advanced content caching systems to reduce retrieval times
- Enhanced CI/CD pipelines and established performance benchmarks

**Amazon** — **Senior Software Design Engineer** *2010 – 2014*

- Designed scalable photo pipeline integrating legacy systems with modern web platforms
- Integrated Amazon Coins into the AppStore
- Built BI platforms using Kinesis, Redshift, and EMR

**Earlier Experience (Condensed)** *1999 – 2010*

**Disney Interactive • Nokia • Dexterra • Microsoft (Volt) • 4MationGeo • Civia • Punch Networks**

Scalable ad platforms, location-based services, enterprise mobile frameworks, GIS systems, automated test harnesses

## SELECTED PROJECTS

**claude-code-search** — Production RAG system for semantic code retrieval. ChromaDB vector storage, sentence-transformers embeddings, benchmarked 12+ HuggingFace models. Maximal Marginal Relevance optimization to prevent redundant code in LLM context. Open source with active users. Evaluating agent orchestration (LangChain) and self-learning indexes (RuVector) for future iterations.

**WA Creel** — Production full-stack web application deployed on Google Cloud Run. RESTful API, automated CI/CD pipeline (Cloud Build), data processing (13 years survey data), interactive visualization (Leaflet, Chart.js). Refactored from 2464-line monolith to modular architecture. Live at **wa-creel.jeremyveleber.com**

## EDUCATION

**Western Washington University** — Bachelor of Arts, Marketing / Business Administration