Anthony J. *Coté*

AI SOLUTIONS ARCHITECT

LOCATION
Edmonton, Alberta, Canada

PHONE
780-231-4825

EMAIL
hello@anthonycote.com

PORTFOLIO
portfolio.anthonycote.com

LINKEDIN
linkedin.com/in/anthonyjcote

21YR BUILDING WEB BASED SYSTEMS	3YR AI/ML DESIGN BUILD CYCLE
70% TOKEN REDUCTION VIA EDGE ML	1.0K+ WEB PROJECTS DELIVERED

CORE STRENGTHS

- AGENTIC AI WORKFLOWS
- INFERENCE COST CONTROL
- CONTEXT ENGINEERING
- RAG SYSTEM DESIGN
- MULTI-AGENT DELEGATION
- PRODUCTION AI SOFTWARE

BUSINESS VALUE

- Increase revenue through acquisition, conversion, delivery, and process automation systems.
- Reduce overhead through context optimization, intelligent routing, edge deployments and LLM substitution.
- Improve operational leverage with internal tooling, analytics and data management.
- De-risk compliance with e2e encryption, SOC2 compliant data handling and PII redaction processes.
- Identify business gaps, translate them into architecture, and build the technical path to close them.

Deliberate Architecture. *Creative Execution.* Measurable Results.

AI Solutions Architect with 21 years of hands-on experience building digital systems — from early web and creative production through to modern AI-enabled software and automation.

Over 3 years in an intensive AI/ML build phase developing production systems for real-world deployment. **Reduced API inference costs by 70%** through dynamic context management, schema compression, edge ML, and cost-aware routing. Trained custom ML models and designed multi-agent workflows, custom tooling, RAG pipelines, model-routing logic, and deployment architecture. Developed and integrated first-party rust and typescript libraries for **ai guardrails, encryption, PII redaction, and SOC 2-aligned data handling.**

Best fit for companies moving fast in uncharted territory. Three-time technical founder with a digital marketing background. A rare mix of technical ability and G2M execution experience. **One part creative director, two parts systems architect, zero templates required.**

PROFESSIONAL EXPERIENCE

AI Infrastructure Architect, R&D

AC92 [03/25 → NOW]

- Developed a custom non-autoregressive language model for air-gapped CPU inference, achieving **sub-20ms response latency** and up to **1,200 t/s** on 8GB consumer hardware.
- Built a vertically integrated software supply chain and **digital workforce** spanning development, deployment, operations, marketing, and commercialization. Enabled full lifecycle ownership and rapid product launches **reducing SaaS costs by 100%.**

AI SaaS Product Architect

GENFIVE [04/23 → 03/25]

- Engineered autonomous LLM orchestration before agentic frameworks were mainstream, using self-awake cron loops, planner-driven execution, and structured context management to overcome context-window limits on early LLM models.
- Built and commercialized a production content publishing system, including a custom licensing and validation API with recurring billing; **exited via Acquire.com**.

AI Automation & Digital Solutions

AIDAMETRIX [04/23 → 05/25]

- Delivered AI systems across dozens of client projects, **owning stakeholder alignment, project scope, and end-to-end execution** from discovery to deployment.
- **Eliminated 100% of wasted proposal meetings** by building an AI-driven quoting system that scoped, priced, and collected payment before any live consultation.

Digital Marketing Specialist

ROHIT GROUP [11/22 → 04/23]

- Managed a **$25K/month digital advertising budget** across paid media, landing pages, and lead generation within a **Canadian Profit 500 environment.**
- Deployed and validated an **automated analytics platform**, providing real-time KPI visibility across channels through comprehensive data auditing.

Digital Solutions Consultant

FREELANCE [06/10 → 11/22]

- **~1,000 web experiences** built, **$2M+ in personal ad spend** deployed, and **$1M+ in annual e-commerce revenue** generated across a 12-year independent practice.

Analyze. *Engineer.* Deploy.

DEV LOGIC

Every system starts with the right question: does this actually need AI? The answer shapes the architecture. If the problem can be solved with deterministic logic, it should be. If it needs language understanding, edge ML runs before cloud inference. If a model is genuinely required, the smallest capable one ships first. The result is modular, system-agnostic code built to deploy across native mobile, desktop, and cloud without re-architecture.

SUPPORTING CREDENTIALS

Education & Professional Development

Software Design and Architecture Specialization
UNIVERSITY OF ALBERTA

Scalable software design, OOD, design patterns, UML, MVC, microservices, SOA, RESTful APIs, and a Java Android capstone evolved into a distributed multi-user system.

Software Product Management Specialization
UNIVERSITY OF ALBERTA

Agile product planning, stakeholder needs, technical specifications, prioritization, and software product execution.

Marketing Strategy Specialization
IE BUSINESS SCHOOL

Market research, positioning, marketing mix, marketing plans, KPIs, budgets, and consumer-centric strategy.

Corporate Strategy
UCL SCHOOL OF MANAGEMENT

Corporate advantage, diversification, divestiture, business strategy, and value-creation decisions.

Digital Leadership and Digital Strategy Execution
DIGITAL MARKETING INSTITUTE

Digital strategy execution, CX design, digital team enablement, and risk-aware marketing leadership.

Responsive Web Design
UNIVERSITY OF LONDON

Interaction design, information architecture, responsive layouts, JavaScript templates, and web usability fundamentals.

LANGUAGES, FRAMEWORKS & RUNTIMES

- Rust
- TypeScript
- Python
- SQL
- React
- Tauri
- Electron
- NestJS
- Next.js
- Astro
- Llama.cpp
- Candle + ORT
- Three.js
- GSAP
- Unity 3D / C#

AI DEVELOPMENT TOOLS

- VS Code + GitHub
- Codex + Claude Code
- Google Stitch + Figma
- DEVMAP (Repo Hygiene)

AI & SYSTEMS ARCHITECTURE

- Inference cost optimization
- Model routing + multi-agent
- ML & LLM orchestration
- RAG / knowledge systems
- Vector databases
- Agentic workflow design
- Tool / function calling
- Prompt & context engineering
- Structured generation
- Human-in-the-loop gates
- Eval, telemetry & debug
- Classical ML integration
- ML training workflows
- Advanced UI prototyping
- API architecture
- Runtime architecture
- Database design
- Local-first systems

DEPLOYMENT & APP TARGETS

- Cloudflare Workers
- Cloudflare Pages
- Serverless architecture
- Static + edge deployment
- Edge AI deployment patterns
- Desktop-native applications
- Mobile-native targets
- Server-based web systems

BUSINESS & GROWTH SYSTEMS

- Solution discovery
- Systems integration planning
- Business process automation
- Client onboarding automation
- Lead qualification
- Proposal automation
- Analytics / KPI systems
- Revenue-focused automation
- Cost / overhead reduction
- Stakeholder translation
- Conversion architecture
- MarTech integration
- Operational leverage design

MODEL PLATFORMS & AI PROVIDERS

- OpenAI API
- Anthropic API
- Google Vertex AI
- Cloudflare AI
- AWS Bedrock
- Azure AI Foundry

$$X_i \therefore Y_i$$
$$Y_i \therefore M$$
$$(X_i, Y_i) \in \Omega(M)$$
$$X_i \neq Y_i$$

Research & Publications: The Law of Instrumental Integrity (Cote, A. 2025) AI SAFETY

A formal four-term structural law for any objective-directed system. Identifies four failure modes — including *terminal logic*, where agent self-continuation displaces the original objective — as a diagnostic extension of Anthropic's published research on agentic misalignment. → anthonycote.com/the-law

Don't Take My Word for It. *Built to Be Experienced...* Not Politely Scrolled Past.

If you're looking for a generic chatbot screenshot with "Innovation" in the caption, this is the wrong portfolio. This is a live collection of immersive 3D interfaces, audio-reactive voice agents, data systems, ML experiments, and agentic RAG demos in a production environment. → portfolio.anthonycote.com