IGOR GANAPOLSKY

Senior AI Systems Engineer (LLM Infrastructure, Cloud, Distributed Systems)

📍 Coral Springs, FL • 📧 iganapolsky@gmail.com • 📱 (201) 639-1534

GitHub: github.com/IgorGanapolsky

LinkedIn: https://www.linkedin.com/in/igor-ganapolsky-859317343/

SUMMARY

Senior AI and Full-Stack Engineer with 15+ years of professional software development experience and 6+ years of significant full-stack development responsibility. Focused on building production AI systems: LLM gateways, agent/tool execution, retrieval-augmented workflows, and cloud-native services on Google Cloud Platform and AWS. Provides technical leadership across architecture, code review, and AI adoption; mentors teams on reliability, cost/latency tradeoffs, and observability. Proven track record building scalable, secure, performant systems aligned with business objectives.

AI/LLM SYSTEMS

• LLM routing & gateways: Tetrate Agent Router Service (TARS), provider fallbacks, cost-aware model selection, prompt/cost controls
• Tools/agents: agentic workflows, tool-using agents, prompt engineering, evaluation & guardrails
• RAG: retrieval-augmented systems, indexing/chunking, relevance tuning, latency/cost tradeoffs
• Cloud AI: Google Cloud Platform (Vertex AI, BigQuery, Cloud Functions, Cloud Build), AWS (Lambda, Bedrock)

CORE COMPETENCIES

• AI Systems / Infra: distributed debugging, reliability, observability (logs/metrics/traces), cost/latency tuning, incident-minded engineering
• L7 traffic & platform: API gateways, HTTP/2 concepts, service mesh fundamentals (Istio/Envoy), production hardening and rollout safety
• Full-Stack Development: React Native (New Architecture, Fabric), Node.js, REST APIs, microservices, cloud-based systems
• Cloud Platforms: GCP (Vertex AI, Dialogflow, BigQuery, Cloud Functions, Cloud Build), AWS (Lambda, Bedrock, S3)
• DevOps & CI/CD: GitHub Actions, GCP Cloud Build, CircleCI, Azure DevOps, Gradle, containerization
• Leadership: architecture reviews, code review, mentoring engineers, clear communication with engineers and stakeholders

PROFESSIONAL EXPERIENCE

Team Lead Mobile Engineer / AI-ML Engineer

Subway Corporate — Miami, FL Jan 2024 – Present

- Led design and delivery of end-to-end AI features from prototype to production, including LLM-backed search, personalized recommendations, and conversational AI assistant serving millions of users monthly

- Architected and deployed Dialogflow CX-based conversational agents integrated with Vertex AI for intent classification and dynamic response generation, reducing customer service load by 35%

- Designed and implemented RAG pipelines combining GPT-4o with app/ordering data to surface personalized menu recommendations, optimizing prompts for relevance, latency, and token cost

- Built secure, scalable APIs and cloud workflows on GCP (Cloud Functions, Cloud Build, BigQuery) to serve LLM features, integrated with existing microservices and CI/CD (GitHub Actions)

- Acted as AI tech lead: defined reference architecture for AI-enabled microservices including auth, rate limiting, logging, and feature flags ensuring secure, observable, and maintainable systems

- Drove best practices for prompt versioning, feature flags for AI, and fallback behavior; mentored engineering team on Windsurf, Cursor IDE, and Claude Code CLI
- Migrated React Native app to New Architecture (Fabric renderer, TurboModules, Bridgeless mode) with Expo SDK 54, achieving 40% faster startup and native-level UI performance across iOS and Android

Mobile Engineer / AI Developer

CNH Industrial — Remote Dec 2022 – Jan 2024

- Built a Gemini-powered GenAI chatbot on Vertex AI for field support, integrating retrieval over manuals/telemetry and workflows for triage; significantly reduced human support requests
- Implemented Dialogflow-based conversational interface for farmer queries, with custom fulfillment webhooks connecting to BigQuery for real-time crop analytics
- Partnered with product and data teams to define prompts, tools, and escalation rules; instrumented usage/quality metrics to continuously improve answer accuracy
- Built real-time farming analytics features in Android app used by 10,000+ farmers with BigQuery integration
- Developed Android UI in Jetpack Compose and managed app state via ViewModel and Flow; increased crash-free sessions to 98%

Mobile Security Engineer (Contract)

Google — Remote Feb 2021 – Nov 2022

- Developed ML- and LLM-assisted malware triage workflows on Vertex AI, improving detection speed and enabling safer app review at Play Store scale
- Built automated pipelines on GCP (Cloud Build, Cloud Functions) for continuous model runs, logging, and performance monitoring across thousands of Android apps
- Contributed to Play Store protection efforts, helping detect and remove 100+ malicious apps through ML-based detection systems on Google Cloud infrastructure

Lead Android Engineer

Abbott Laboratories — Alameda, CA Feb 2020 – Jan 2021

- Led Android team on FreeStyle Libre app, used by millions of users for glucose monitoring; maintained HIPAA compliance
- Re-architected Android app with MVVM + Kotlin Coroutines for enhanced reliability and scalability
- Integrated BLE + MLKit for sensor connectivity and glucose prediction; optimized for battery and performance

Senior Android Developer

Crestron Electronics — Rockleigh, NJ Dec 2013 – Dec 2019

- Shipped Crestron Home Android app using Kotlin and MVVM, serving 10,000+ users with real-time IoT automation
- Improved communication latency by 20% via custom IoT stack; rolled out CI/CD using GitHub Actions

Senior Software Engineer (Java/Android)

KPMG LLP — Montvale, NJ Aug 2005 – Aug 2012

- Engineered Java-based enterprise solutions; acted as Scrum Master optimizing dev velocity by 20%

- Implemented RESTful services and CI/CD pipelines; pioneered internal Android POCs

SELECTED AI PROJECTS

• Autonomous AI trading system (github.com/IgorGanapolsky/trading): multi-model LLM gateway with Tetrate Agent Router Service (TARS), cost-aware routing and provider fallbacks, and self-healing CI for continuous reliability
• Conversational AI ordering assistant for Subway (Dialogflow CX, Vertex AI, GPT-4o, RAG, GCP) – production deployment
• Field-support GenAI chatbot for CNH (Gemini, Vertex AI, Dialogflow, BigQuery) – reduced support volume 40%
• Malware detection and triage system at Google (TensorFlow, Vertex AI, GCP Cloud Build) – Play Store scale ML
• HIPAA-compliant glucose prediction app with BLE + TensorFlow Lite integration at Abbott