

Gunveer Sandhu

Applied AI Product Engineer | LLM Systems, On-Device ML & Production Engineering

Bikaner, India • Open to relocation

+91 89529 26407 • gunveer.sandhu11@gmail.com • [LinkedIn](#)

SUMMARY

Applied AI product engineer with nearly five years of production engineering experience building AI-native products end to end. Independently built 0→1 systems spanning model adaptation and evaluation, low-latency on-device inference, structured multimodal LLM workflows, backend infrastructure and customer-facing applications. Earlier work includes products used by millions and complex production systems across Android, iOS, macOS, Flutter and web—strongest when owning an ambiguous problem from technical decisions through production release.

PROFESSIONAL EXPERIENCE

Founder & Product Engineer | Applied AI & Consumer Products

Nov 2025–Present

SpellType | On-Device AI Autocomplete for macOS

Generative AI • Machine Learning • Large Language Models (LLMs) • Production LLM Systems • LLM Evaluation (Evals) • Model Fine-Tuning (LoRA) • Prompt Engineering • On-Device LLM Inference • Inference Optimization • Quantization • MLX/MLX-LM • KV-Cache Optimization • Token Healing • Swift Concurrency

- **Built and shipped a production on-device LLM autocomplete system for macOS**, running OLMo 2 1B locally on 8 GB Apple Silicon with **sub-200 ms median suggestion latency** across 10+ writing apps; owned production delivery with verified model distribution, Developer ID signing, notarization and signed auto-updates.
- **Fine-tuned OLMo 2 1B with LoRA across all 16 transformer layers** using 1,680 curated training examples; in a blinded 200-case evaluation, the deployed 8-bit model was **preferred roughly 3:1** over the unadapted 8-bit baseline.
- **Engineered a low-latency MLX inference runtime for continuous typing**, using persistent KV-cache reuse, token healing, idle prewarming and Swift Concurrency-based cancellation/preemption to minimize redundant prefill work and prevent stale generations during rapid user input.
- **Expanding SpellType toward business-context autocomplete**, designing Retrieval-Augmented Generation (RAG) workflows using embeddings and vector databases for enterprise knowledge retrieval, grounding LLM suggestions in approved company content and workflow context with context engineering and guardrails.

Cosmicly | Multimodal AI Product • Production LLM Workflows

Multimodal AI • LLM Applications • Agentic AI • AI Workflows • LLM APIs • Prompt Engineering • Structured Outputs • Context Engineering • LLM Orchestration • Model Routing • Guardrails • Retries & Fallbacks • Observability • Google Gemini • Vertex AI • TypeScript • PostgreSQL • Supabase • Serverless Edge Functions • Flutter

- **Built Cosmicly end to end as a production cross-platform AI product**, spanning Flutter iOS/Android clients, Supabase/PostgreSQL backend services and Vertex AI/Gemini workflows, with authentication, realtime data, subscriptions, push notifications, localization and production delivery.
- **Architected a hybrid deterministic + LLM system**, building and testing an astronomy engine for planetary positions, chart and timing calculations, then layering **nine schema-constrained Gemini text and vision workflows** on top to keep deterministic computation separate from probabilistic generation.
- **Built the production LLM reliability layer** with workload-specific routing across Gemini Flash-Lite, Flash and Pro, region-aware fallbacks, bounded retries/timeouts, truncation handling, malformed-JSON recovery, cached authentication and **per-feature token/cost observability**.
- **Designed the production Supabase/PostgreSQL backend** with Row-Level Security, realtime data flows, database functions/triggers, version-controlled migrations and TypeScript/Deno Edge Functions; built idempotent subscription webhooks, automated backend/client tests and **staging/production CI/CD**.

NoRing | Android call & messaging protection • Google Play • 30 languages

Kotlin • Jetpack Compose • MVVM • StateFlow • Coroutines/Flow • WorkManager • Room/SQLite • Android Telecom/Telephony • SMS/MMS • Multi-SIM • Play Integrity • Crashlytics • Google Play • Testing • CI/CD

- **Designed the on-device decision engine for Android's 5-second call-screening window**, with explicit precedence across user rules, contacts, outgoing-call history and learned exact-number/pattern behaviour; kept provider IPC and non-critical work off the screening path, with fail-safe handling and a 5-minute repeat-caller safeguard.
- **Built the Compose application around reactive MVVM and repository-driven state**, using StateFlow, SavedStateHandle and lifecycle-aware coroutines to keep UI, Room and Android system-provider state consistent across pagination, process recreation, default-SMS role changes, retries and partial failures while supporting full SMS/MMS and multi-SIM workflows.
- **Built a deterministic on-device SMS protection engine backed by Room**, classifying messages across 10 semantic categories using content, sender type, contacts and conversation history; applied explicit sender/conversation rules before automatic filtering, keeping OTPs, transactions and other important messages visible while separating promotions and suspicious messages without sending message content off-device.

Olly | 0→1 fertility & ART product · iOS

Swift · SwiftUI · UIKit · Swift Concurrency · REST APIs · Apple Health · Localization · App Store

- **Owned 0→1 iOS delivery** for Olly, a production ART/IVF companion promoted at CES 2024; as one of two frontend engineers in a five-person international product team, built the native SwiftUI/UIKit client through App Store launch and ongoing production delivery.
- **Built complex healthcare workflows** for medication and appointment tracking, fertility journeys, medical histories, blood tests, questionnaires, document upload/preview/sharing, onboarding and alerts—integrating API-driven state across a sensitive, multi-step domain.
- **Evolved the client through sustained releases** with two backend engineers and a technical lead, using Swift Concurrency, reusable UI architecture and REST integrations; added Apple Health-connected fertility tracking and localization across English, French and Spanish.

GreyMonk | AI-powered engineering intelligence product

TypeScript · React · Frontend Architecture · REST APIs · Data-heavy UI · AI-generated Insights

- **Built substantial TypeScript/React frontend** for an AI-powered engineering intelligence platform, turning code, pull-request, delivery and team-performance data into useful product experiences for engineering leaders.
- **Developed data-heavy product flows** for engineering metrics, AI-generated insights, team/project views and delivery signals, managing asynchronous REST API state and translating complex backend data into clear interactive interfaces.
- **Partnered with backend engineers** on API contracts, data models and product behavior, contributing to technical decisions about how engineering signals and AI-derived outputs were represented and consumed in the frontend.

Mosaic | production POS · Flutter

Flutter · Dart · Android Runtime · WebView · REST APIs · Firebase · Performance Profiling · ANR/Crash Debugging · Production Reliability

- **Stabilized and modernized** a production Flutter POS application, resolving recurring ANRs, crashes, runtime errors, slow interactions and reliability failures across critical workflows.
- **Eliminated several ANR-prone WebView paths** by rebuilding them as native Flutter flows, while progressively refactoring brittle legacy code and improving application structure without disrupting ongoing product delivery.
- **Added Firebase production instrumentation and analytics**, using real-world feedback alongside profiling and debugging to improve performance, responsiveness and reliability.

Mobile Application Developer (Android) | RailYatri (Stelling Technologies)

Sep 2021–Sep 2023

RailYatri & IntrCity SmartBus | Consumer Android Products Used by Millions

Kotlin · Java · Android SDK · Views/XML · Coroutines/Flow · Retrofit/OkHttp · Room/SQLite · MVVM/MVI/Clean Architecture · Modular Gradle · Testing · Profiling · Observability · CI/CD

- **Built and maintained five production Android applications**, including two consumer products used by millions, working end to end across architecture, UI, networking, local persistence, asynchronous state, testing, debugging and releases.
- **Modernized substantial legacy Java code to Kotlin** and strengthened production architecture with MVVM/MVI patterns, Coroutines/Flow, Room/SQLite and modular Gradle; raised crash-free usage from 92% to 98% through profiling, observability and reliability improvements.
- **Helped turn a failed government security audit into a passing reassessment within 10 days** by remediating obfuscation, secure local data, production configuration, component exposure, dependency risk, tamper/root detection and device integrity.
- **Led two developers** while remaining hands-on across architecture, implementation, debugging, code quality and production releases.

EDUCATION**Bachelor of Technology in Computer Science Engineering**

2017–2021

Jaypee Institute of Information Technology (JIIT), Noida

LANGUAGES

English — Near-native proficiency