

YOONBO CHO

AI Engineer · Applied LLM Systems

Elkins Park, PA · +1 718 541 6008 · yoonbocho1@gmail.com · yoonbo1.github.io · github.com/yoonbo1 · linkedin.com/in/yoonbo-cho-9780301b9
· Open to remote or relocation to NYC, Chicago, or WA

WORK EXPERIENCE

Founder & AI Engineer — Yoonibo LLC

Jul 2022 – Present

tickatap — invitation-first going-out planner · **tickatap.com** · live, ships daily

- Shipped **tickatap.com** solo and operate it live: an invitation-first going-out planner where a host picks the day, each friend taps what they're into from a full-screen board of real events and places, and the plan builds itself the moment the last person finishes.
- Built the board's daily **event-curation agent** on a two-tier model split: high-volume page extraction on a local 14B model chosen by a measured head-to-head, and one frontier-model curation call per city. A **193-source registry** spanning 40+ US cities, every URL probed live before enrollment; cron runs carry locks and freshness guards, and a missing dependency reports an honest skip rather than fabricated events.
- Built **Tap**, a booking agent, on a hard safety invariant: purchase execution required an active mandate enforced server-side in the tool executor, never the prompt, verified by a 25-case adversarial suite at 100%. Retired it at 13.3% task success against a 90% bar, and kept the invariants.
- Built the platform as a **10-workspace TypeScript monorepo** (Fastify, Next 15, Expo, Postgres) holding **3,682 tests passing, 0 failing**, the entire suite running with no database, no network, and no vendor keys; every commerce integration sits behind a provider interface with a deterministic mock, so the full checkout path runs and tests offline.
- Built a short-form video studio grounded in live product data: the model writes each script choosing only from real candidate places, every pick validated back by name, so it cannot invent a stop; renders headlessly to vertical video, and a human approves every post.

Everblue — Yoonibo LLC

Passive AI journaling iOS app · **everblue.me** · live on the App Store, launched Jun 2026

- Built an **LLM eval harness** that runs the real production pipeline on synthetic day-payloads and grades against the production rubric (voice lint, specificity, hallucination). Ship bars of 92%+ generation quality and a 100% no-fabrication gate block every prompt change.
- Engineered an **on-device hybrid-retrieval RAG pipeline**: MiniLM ONNX embeddings, BM25, exponential recency decay, and MMR re-ranking, grounding an AI companion in months of personal context. Deliberately no vector database: the per-user corpus doesn't warrant one.
- Built a **multi-provider LLM backend** (Django) as a provider abstraction with byte-identical prompt fallback, routing by feature, with retry-and-fallback orchestration, cache-stable prompt prefixes, per-device daily request budgets, and PII redaction on every request.
- Shipped fully on-device photo captioning (~500MB FP16 **Florence-2** Core ML, custom Expo config plugin, Swift bridge) and debugged its jetsam crashes on 4GB iPhones; then reversed the design when the ~1.1GB bundle stopped earning its cost, replacing it with a gated backend service.

Senior Developer — A View From My Seat

Nov 2024 – Jul 2026 · Philadelphia, PA

Sports/venue platform · **1.2M monthly active users**

- Owned native iOS (Swift) and Android (Kotlin) development end-to-end: spec, architecture, implementation, QA, and release, shipping **80+ updates** at feature parity across platforms.
- Shipped **vision-based moderation** for user-submitted seat-view photos, auto-flagging or clearing each image for publication, and backfilled structured metadata like view obstructions across the photo corpus.
- Built a **venue-scouting agent** that runs daily, diffs live venue data against stored snapshots, and alerts only on real changes, replacing manual monitoring.

Software Developer — Fabspeed Motorsport

Jan 2022 – Jul 2022 · Philadelphia, PA

- Launched **wholesale.fabspeed.com**, taking the site from under-construction to production-ready, and built reusable React tools for bulk product updates.
- Built **Python scrapers** extracting Google and Yelp customer reviews at scale.

CERTIFICATIONS

TensorFlow Developer Certificate — image classification, NLP, and time-series forecasting