
Ahmed Afridee

Dhaka, Bangladesh • +8801910027738 • a.afridee47@gmail.com • [linkedin.com/in/ahmed-afridee](https://www.linkedin.com/in/ahmed-afridee) • <https://github.com/Afridee>

Flutter Developer | AI Engineer

Flutter Developer and AI Engineer with 5+ years shipping production Android and iOS apps in Dart. Built and maintained offline-first, high-volume mobile systems serving 4500+ field agents, and integrated on-device AI pipelines using LangChain, LangGraph, flutter_gemma, and RAG — from architecture through App Store deployment. Published in Towards AI (2026) on on-device LLMs in Flutter.

WORK EXPERIENCE

Manush tech • Singapore

04/2025 – Present

Flutter Developer • Full-time

- Eliminated order entry failures in zero-connectivity zones for 4500+ field agents by building an offline-first capture layer with Drift — orders now complete successfully regardless of network state
- Reduced sync errors by 99% and eliminated data loss across 50,000+ daily reconnection events by designing a bidirectional sync engine with deterministic conflict-resolution logic, co-owning API contracts with backend engineers
- Architected a Drift-backed local schema supporting 15+ entity types and 50K+ daily writes, enabling complete offline parity with the server-side data model — zero feature degradation without connectivity

Elements Group • Sydney, NSW

02/2020 – 04/2025

Flutter Developer • Full-time

- Sole developer on a live employee management app, used by 257 employees across client organisations
- Eliminated fraudulent attendance entries entirely — achieving 0 incidents post-launch and removing the need for manual supervisor sign-off — by building a GPS-verified clock-in system using Geolocator
- Developed a dynamic checklist engine allowing managers to configure and assign day-specific task forms to field employees
- Delivered real-time push notifications (FCM) for shift reminders, task alerts, and broadcasts to 257 employees — achieving 98%+ delivery reliability across iOS and Android over 5 years
- Architected the app with GetX for state management and Dio for REST API communication, maintaining the codebase solo for 5 years across 10+ releases
- Configured Codemagic CI/CD pipeline for automated builds and TestFlight/App Store distribution, maintaining zero manual deployment overhead across 5 years of solo development
- Wrote unit and widget tests covering core business logic including geo-clock-in validation and checklist state management

PUBLICATIONS

I Crammed RAG, a Vector Database, and a Gemma LLM into a Mobile App. Here's What Happened.

05/2026

Towards AI

<https://pub.towardsai.net/i-crammed-rag-a-vector-database-and-a-gemma-llm-into-a-mobile-app-heres-what-happened-6e1a270e6d44>

AI ENGINEERING PROJECTS

Karo- Building an AI Agent That Knows Your Company

- Built an AI agent that ingests internal knowledge (API docs, business rules, procedures), enabling team members to query live data in plain English — no SQL, Postman, or engineering tickets needed
- Added an eval harness for regression testing retrieval accuracy and agent behavior across prompt/schema changes; fully containerized with Docker Compose
- Implemented a hybrid RAG + API-calling architecture using LangGraph/LangChain, with pgvector for semantic retrieval and a dynamic HTTP tool supporting per-request auth injection
- Persisted multi-turn conversation history in PostgreSQL via LangGraph's PostgresSaver; built chat UI in Chainlit with text and voice input (server-side transcription via faster-whisper)
- **Stack:** Python · LangGraph · LangChain · OpenAI GPT-4o · pgvector · PostgreSQL · Chainlit · Docker

Demo : <https://youtu.be/1x4H6KBWVS8?list=PLQJcnbfceQTTFeGbseG-k2VfKO4GEaHo1>

Github repo: <https://github.com/Afridee/Karo>

Smart Notes

- Achieved fully private, offline AI inference on-device — zero cloud calls, zero API keys — by integrating Gemma 4 E4B IT (~4.3 GB) and EmbeddingGemma via flutter_gemma for generation and 768-dimensional text embeddings
- Implemented hybrid retrieval combining dense cosine search (HNSW via ObjectBox) and BM25 keyword scoring, fused with weighted-sum normalization ($0.7 \times \text{dense} + 0.3 \times \text{BM25}$) and MMR re-ranking for result diversity
- Designed a semantic graph view using pairwise cosine similarity across note mean vectors, surfacing topical connections without tags or folders
- Managed a fixed 2,048-token context window with dynamic token-aware prompt trimming across system instruction, retrieved chunks, and query

Github repo: https://github.com/Afridee/smart_notes

SKILLS

- **Mobile & Flutter:** Flutter, Android, iOS, Dart, flutter_gemma, GetX, Provider, Riverpod, BLoC, Dio, Drift, SQLite, Geolocator, Push Notifications, Unit Testing, Widget Testing, Responsive & Adaptive UI, Codemagic CI/CD, Firebase, Git, GitHub
- **AI & Backend:** Python, LangChain, LangGraph, OpenAI GPT-4o, Gemini API, RAG, pgvector, PostgreSQL, Vector Search, Embeddings, ObjectBox, BM25, on-device ML, Multimodal, Chainlit, FastAPI, Docker, REST APIs, faster-whisper, prompt engineering, agent architecture, hybrid retrieval pipeline
- **Tools & Workflow:** Agile, Scrum, CI/CD, Docker, Codemagic

EDUCATION

B.Sc. in Computer Science and Engineering

North South University

01/2018 – 08/2022