

PROFILE

Every product I have shipped started as something I wanted in my own life: how I plan my days, how I learn a language, a walking game, the seating for a meetup I go to. Building them taught me the whole craft — models, backend, app and deployment — and eventually I built the platform they all needed. **I like turning an idea into something real, then keeping it alive.** Now I want to do it inside a team, on something more people depend on.

01 EXPERIENCE

Yumi Nexus Founder & AI engineer · self-directed, pre-revenue Mar 2026–Present

A production-deployed, multi-tenant generative-AI platform that turns natural-language intent into secure, observable action. Built and operated solo in three layers with **strictly one-way dependencies**: no layer sees the one above it, and the core ships as the versioned package the platform installs.

L1
Agent core
[Source code ↗](#)
[PyPI ↗](#)
Apache-2.0

AGENT LOOP A streaming turn loop over **provider-native function calling**, with persistent memory across turns. Three hosted providers — Claude, Gemini, OpenAI — plus local Ollama and fastembed behind one interface, sharing error classification and a warm-up / pull / shutdown lifecycle.

TOOL EXPOSURE Three registration modes, chosen so the agent **gathers before it speaks**: **autorun** runs ahead of generation and injects only its result; **pinned** stays in the list; **dynamic** is retrieved against the query.

PROMPTS Layered prompt composition under a semantic version **plus a content hash**, so wording edited without a version bump is caught automatically.

MEMORY Two stores by design: **SQLite as the source of truth**, **LanceDB for vector search**, reconciled both ways; during an index rebuild concurrent writes commit to SQLite only, so they cannot duplicate rows.

SHIPPING **Function-calling bindings for 11 targets**, a CLI and a turn inspector; **699 tests** with mocked providers, CI on Python 3.10–3.13, released to PyPI by OIDC Trusted Publishing.

L2
Multi-tenant platform
yumi-enterprise
closed source

TENANCY Per-tenant isolation over PostgreSQL, with Argon2 hashing, OAuth, linked chat accounts and auth rate limiting.

METERING Quotas measured in **turns, tokens and dollars** across 21 admin endpoints and a React console.

OBSERVABILITY & MONITORING A console for finding faults: an aggregate view of how requests were routed to tools, and the exact execution trace behind any single turn.

GOVERNANCE & PRIVACY Two tiers enforced in SQL, not policy: history views select only message length, never content, and revealing a turn **writes the audit record first, failing 503 if that write fails**.

L3
Production & operations
[Yumi homepage ↗](#)
[Sign-in page ↗](#)
GCP

DEPLOYMENT GCP Compute Engine behind Caddy/TLS with Docker Compose, health checks, traces and metrics; serves yumi.nexus, the identity site and a live agent demo.

IDENTITY One Firebase-backed sign-in serves the site, the agent and every edge app, with **cross-app handoff**; authorisation lives in hand-written Firestore rules.

INFRASTRUCTURE AS CODE The production GCP project is managed in **Terraform** on a locked remote backend — the running estate imported and held to a clean **“No changes”** plan first, then five reviewed changes, including **SSH moved to IAP-only**.

CONTINUOUS DELIVERY A push deploys itself over **GitHub OIDC into Workload Identity Federation — no stored cloud credential**: three repos rebuild and health-check behind a host lock, **any failure rolls all three back**, and the public hostname is then re-verified.

CONTINUITY Scripted backups with retention, and a **restore drill actually executed**: snapshot to a fresh disk, restored onto a network-isolated VM, then 21 tables diffed.

EDUCATION

Master of Artificial Intelligence

University of Auckland
Thesis: CiteSeek · submitted, result pending
Feb 2025 – Jul 2026

BEng, Communication Engineering

Tianjin University, China
Thesis: Ship RL
Sep 2019 – Jul 2023

WORK ELIGIBILITY

AVAILABLE Eligible to apply for a 3-year open NZ Post Study Work Visa once Master’s completion is confirmed (~Oct 2026); full-time from ~Nov 2026, subject to approval. No sponsorship required.

UNTIL THEN Student visa: 25 hrs/week, full-time in university vacations.

CORE SKILLS

Generative AI & agents

AGENTS Agent loops, tool calling, MCP, persistent memory

RETRIEVAL RAG, embeddings, hybrid search, vector DBs

MODELS Gemini, OpenAI, Claude

Machine learning

FRAMEWORK PyTorch

RL TD3, SAC, PPO

CV YOLOv5, SiamRPN

NLP LSTM, Transformers, attention

Backend & data

CORE Python, SQL, async

APIS FastAPI, REST, NDJSON, WebSocket

DATA PostgreSQL, SQLite, Firestore (NoSQL), LanceDB, Snowflake

PLATFORM Auth, multi-tenancy, quotas, usage metering

App development

MOBILE Flutter / Dart, SwiftUI / ARKit

WEB TypeScript, React, Streamlit

Cloud & deployment

GCP Compute Engine, Cloud Storage, Secret Manager, Firebase, IAP

IAC Terraform, Docker Compose, Caddy / TLS, backups

Engineering practice

CI/CD GitHub Actions, pytest + mocking, Ruff, OIDC keyless deploy

PUBLISHING PyPI, pub.dev · tracing, metrics

LANGUAGES

ENGLISH Professional working proficiency

MANDARIN Native

JAPANESE JLPT N2

02 SHIPPED SOFTWARE

Mugi ↗ 2026	CROSS-PLATFORM DELIVERY One Flutter codebase shipping to iOS, Android, macOS and web, localised into ten languages. ARCHITECTURE An adapter interface lets the same app run on Firebase or my own API, chosen at bootstrap rather than baked into the screens. SPEC IMPLEMENTATION Recurring tasks expand through a 371-line RFC 5545 implementation rather than a package, with its own tests. Flutter · Firebase · iOS / Android / macOS / web
Memori ↗ 2025–2026	ALGORITHM DESIGN Urgency is $\log(1 + \text{days}) \times (6 - \text{mastery})$, sampled through a softmax with a reinforcement queue and recent-pick avoidance — so there is no “due” pile to punish a missed day. PLATFORM MIGRATION Shipped first on SwiftUI, then moved the whole app to Flutter: the 452-line scheduler was ported in full and pinned by tests. APPLIED LLM Gemini generates structured vocabulary cards across ten languages. Flutter · Gemini · Cloud Functions · adaptive scheduling
Mugi × Memori RUN ON YUMI	Mugi and Memori are not standalone apps that happen to call an API — each runs as an edge of the platform above. A user can drive either one by talking to Yumi, and before every reply Yumi pulls a live snapshot from both: today’s mood, what is due, which words are weak. The answer is grounded in the user’s actual state rather than asked for. Each snapshot is guardrailed — capped, cached for 60 seconds and marked reference-only — so it informs the reply without being recited back.
SaKi ↗ 2025–2026	SENSOR PROCESSING Camera poses pass an exponential filter with clamped gain, jump rejection, and a stabilisation gate that waits for consecutive coherent frames before sampling resumes after relocalisation. REAL-TIME AR A recorded path is rebuilt as geometry the player walks back: warning past 35 cm, failing only on sustained deviation beyond 1.1 m, both tunable. RELEASE ENGINEERING Built in SwiftUI and ARKit, through App Store review and live. SwiftUI · ARKit · haptics · spatial audio
Meetup Seater ↗ 2025–Present	COMBINATORIAL OPTIMISATION Seating is a weighted greedy assignment: having met before costs 1, sharing last round’s table costs 3, and each person joins the table with the smallest increase in repeats. CONSTRAINT MODELLING Table count and language balance are solved together, so English and Japanese speakers stay mixed. OPERATIONS Run every week for ~30 people by an organiser who is not me, from phones, with live check-in. Streamlit · Firestore · live phone views

03 RESEARCH FOUNDATIONS

CiteSeek ↗ Master’s thesis 2026	INFORMATION RETRIEVAL Searches scholarly APIs and follows citation graphs to reach what a keyword search misses, then verifies evidence at passage level so a citation must be supported by a specific span. DATA PIPELINE PDF and HTML full text parsed, chunked, embedded and deduplicated before anything is ranked. OFFLINE EVALUATION A hand-built 28-claim golden set run across nine configurations against a BM25 baseline. Citation expansion lifted the retrieval ceiling from 22/28 to 28/28 claims; the full pipeline reached Hit@5 0.964 · R@5 0.911 · MRR 0.837, against 0.321 · 0.298 · 0.248 for BM25. ABLATION Varying the embedding and judge models showed the structural citation signal, not model choice, carried the result — ranking by how many seed papers cite a candidate reached Hit@5 0.893 before any LLM judging at all. SYSTEM INTERFACES Driven from a FastAPI workbench, a CLI and an eight-tool MCP server — the same interface used to write claims, evidence and provenance into a supervisor’s agentic system without modifying its source. Python · agentic RAG · scholarly APIs · embeddings · MCP server
Ship RL BEng thesis 2023	PHYSICAL MODELLING A 4-DOF MMG simulator parameterised from a real 175 m container ship, with the inertia and actuation lag of a real hull, built before any learning so the results would mean something. REINFORCEMENT LEARNING Soft Actor-Critic implemented in PyTorch from the paper rather than pulled from a library, trained on the two tasks that dominate real navigation — path following and collision avoidance. REWARD DESIGN COLREGs-derived shaping reached 95% collision avoidance after ~4,000 episodes, against 16% for a like-for-like sparse-reward baseline. PyTorch · SAC · MATLAB · pygame · physics simulation