RISHABH RAY

AI Engineer | Applied AI | LLM Systems | Machine Learning
Email: hi@rishabhray.me | 0481 334 132 | rishabhray.me | linkedin.com/in/rishabhrayy | github.com/rishabhrayy
Melbourne, VIC | Temporary Graduate Visa (Subclass 485)

PROFESSIONAL SUMMARY

AI engineer who builds LLM features and RAG systems that keep working when the model doesn't. I've shipped three Llama 3.3 features in an expo-winning product and a red-teamed RAG assistant that runs live on my site, each measured rather than assumed. Comfortable across the stack - Python, TypeScript, PyTorch, PostgreSQL, AWS. Currently mentor undergraduate software teams and set AI-usage standards for a unit at Monash University.

EXPERIENCE

Sessional Teaching Associate - Monash University, Faculty of Information Technology

Clayton, VIC | Feb 2026 – Present · Technical Mentor, FIT3047 Industry Experience (S2 2026)

- Mentor 13 undergraduate student teams (78 students) through a full-semester, industry-sponsored software project, running weekly studio sessions and supervising delivery against client milestones.
- Set and enforce the unit's standards for AI-assisted development, defining where an assistant may operate - noticeably speeding up review turnaround.

Brand Operations Manager - Say It With A Pin (2626 Creative Studio)

Delhi, India | Oct 2023 – Jun 2024

- Built and automated internal operations tooling (Zapier, n8n) integrating Shopify, Stripe and Mailchimp, saving a full day or more of manual coordination each week.
- Owned B2B client delivery for accounts including BMW, Mini Cooper (7,500-unit hand-painted order delivered end-to-end within a one-month turnaround), and Third Wave Coffee.

PROJECTS

NeighbourFit - Melbourne Suburb Recommendation Platform · Winner, Monash PG Industry Experience Expo (S1 2026)

Mar – Jun 2026 | Six-person team, FIT5120 Industry Experience Studio, under an APA Group industry mentor | Vue 3, Flask, AWS Lambda, API Gateway, PostgreSQL on RDS, Groq (Llama 3.3 70B), AWS Amplify

- The problem: choosing where to live in Melbourne meant manually cross-referencing transport, safety, amenities and demographics across separate sites, with no single tool weighing any of it against a person's own priorities.
- Designed and shipped a hybrid AI recommender on Llama 3.3 70B, combining three LLM features (natural-language preference extraction, persona-aware suburb summaries, and qualitative re-ranking of a deterministic top-15 shortlist to three) with enforced JSON schemas, graceful degradation, and visible model reasoning, so the product kept working, and stayed explainable, even when the model didn't.
- Evaluated the voice-to-preferences feature against the live endpoint on a 12-utterance test set: 100% weight accuracy (12/12) and 83% persona accuracy (10/12), with both misses traced to the same ambiguous phrasing.
- Built the weighted scoring engine and the REST API behind it, using UNION ALL queries across spatial point-of-interest and transport-stop tables, scaled to 290+ Melbourne suburbs - cutting a typical evening of manual research down to about 30 minutes.
- Ran fortnightly reviews and requirement-gathering sessions with the team's APA Group industry mentor, translating ambiguous, evolving asks into a scoped three-iteration delivery plan that shipped all three iterations on schedule.

Ask about my work and Workplace Q&A - RAG assistants, live

TypeScript, Postgres + pgvector, Gemini embeddings, Groq · rishabhray.me/projects/ask

- Built a retrieval-augmented assistant that answers questions about my work from my site's published pages only: hybrid BM25 and embedding search, cited answers, and a four-model fallback chain that ends in an extractive answer, so it keeps answering when every model is down.
- Red-teamed it with 75 attacks across 13 categories drawn from published jailbreak research; hardened a pre-model guard from stopping 10.7% to 64% of attacks with no real questions blocked, fixed the one end-to-end success, and now run all 75 against the live models weekly in CI (0 of 75 succeed).
- Reused the engine on the Fair Work Act 2009 and compared six retrieval methods on a 101-question test set with auto-verified answers: embeddings plus an LLM reranker reached 100% hit@5 (MRR 0.967) against 82.2% for BM25, and hybrid search, the better choice on my site, lost to embeddings alone on legal text.

Outfit Picker - solo, shipped side project

React, Vite, IndexedDB, PWA · github.com/rishabhrayy/outfit-picker

- Designed and shipped a solo local-first PWA where AI providers (OpenAI, Anthropic, Gemini, Groq) are swapped at runtime behind one interface, each backed by a four-tier fallback ladder for structured output so a provider's quirks degrade gracefully instead of breaking the app.
- Handled the integration details that make a bring-your-own-key app actually reliable: detecting a key pasted for the wrong provider, working around a provider's reasoning-token budget, and adding a required browser-access header automatically.
- Kept everything client-side by design: photos, wardrobe records and provider keys live in IndexedDB and localStorage, with no backend or account ever seeing them.

AI Systems Portfolio - from first principles

Python, NumPy, PyTorch · github.com/rishabhrayy/monash-projects

- Trained four independent DQN agents from scratch (experience replay, Polyak-averaged target networks, a difficulty-graded curriculum), and built feedforward networks directly on PyTorch tensors, implementing SGD, momentum and AdaGrad by hand with exact ELU/GELU derivatives for backpropagation.
- Built a hybrid agent combining PDDL-inspired symbolic reasoning, a particle filter with Bayesian resampling for hidden-opponent tracking, and risk-sensitive A* navigation.
- Implemented KNN regression from scratch with a KDTree, plus L-fold cross-validation, bias-variance analysis and one-standard-error model selection.

RAY/OS - personal agent system, running daily

Claude, Obsidian, Python, Flask, MCP · rishabhray.me/#rayos

- A personal agent system on Claude: a structured, auditable memory, skills that encode how I work, scheduled agents and tools over MCP, with a human in the loop for anything that leaves it.

KEY SKILLS

Languages: Python, SQL, TypeScript, JavaScript, PHP, C++

AI & ML / LLM engineering: PyTorch, scikit-learn, NumPy, Pandas; Groq (Llama 3.3 70B, gpt-oss), Gemini, OpenAI and Claude (Anthropic) APIs; RAG (hybrid BM25 and embedding search, pgvector, LLM rerankers), retrieval evaluation (hit@k, MRR), LLM red-teaming and prompt-injection defence; prompt engineering, structured JSON output enforcement, agentic workflows and multi-step LLM orchestration, graceful degradation and fallback design; DQN and multi-agent RL, KNN, Bayesian networks, particle filters, A* search, constraint programming (MiniZinc)

Backend & cloud: Flask, REST API design, AWS (Lambda, API Gateway, RDS, S3, EC2, Amplify), Vercel, PostgreSQL with pgvector (Neon), Upstash Redis, MongoDB, Docker, Git, GitHub Actions CI, JWT and bcrypt auth

Frontend: Vue 3 (Composition API), Astro, Vite, Axios, Leaflet.js

Communication & stakeholder engagement: Requirement-gathering and stakeholder translation, code review and mentoring, Agile/Scrum, cross-validation and reproducible experiment design

EDUCATION

Master of Artificial Intelligence - Monash University

Melbourne, VIC | Jul 2024 – Jun 2026

- Coursework: Deep Learning, Machine Learning, Multi-Agent Systems, Discrete Optimisation.