

Adith Kattuparambil Biji

☎ +353 899 707 520 ✉ adithkb@gmail.com [in linkedin.com/in/adithkb](https://www.linkedin.com/in/adithkb) github.com/AdithKB adithkb.dev
Dublin, Ireland

Applied AI Engineer | LLM Applications, AI Evaluation & Production Systems

SUMMARY

Applied AI Engineer building production LLM applications, agentic pipelines, and AI evaluation systems — currently completing an MSc in Artificial Intelligence at NCI Dublin (Academic Merit Scholarship) with 5+ years of engineering experience across capital markets automation and startup product development. Projects include a deployed LLM triage engine (Lette AI), a two-stage agentic PKM pipeline (Nexus v4.3), and a co-authored medical LLM reliability study across 6,000 inferences. Production validation background maps directly to AI reliability work — catching system failures before they reach production is the discipline that transfers most cleanly to evaluating non-deterministic AI behaviour.

TECHNICAL SKILLS

AI/LLM: LangChain, Gemini API, Claude API, OpenAI API, Structured Outputs (Tool-calling), Pydantic v2, Prompt Engineering, MCP Server Development

AI Evaluation: Fleiss' κ , Krippendorff's α , Bootstrap Confidence Intervals, Multilingual LLM Evaluation, Constrained Decoding

ML/NLP: scikit-learn, XLM-RoBERTa (fine-tuning), SVM, XGBoost, TF-IDF, SMOTE, Online Learning (SGDClassifier)

Backend/Infra: Python, FastAPI, Docker, DigitalOcean, Local Inference (MLX/Ollama), PostgreSQL, Supabase, REST APIs

Prior Stack: Java, Selenium, TestNG, CI/CD, Git/GitHub, SQL

AI PROJECTS

Lette AI — Autonomous LLM Triage Engine

[Hackathon] Give(a)Go Dublin, Mar 2026

LangChain, FastAPI, Pydantic v2, React

- Constrained LLM to return a validated 23-field Pydantic model via tool-calling — typed fields and constrained enums ensure valid outputs without regex parsing or post-processing
- Implemented adjustable autonomy threshold (default 0.85) to draft or auto-route decisions; resolves routine inquiries via property knowledge base injection; escalates emergencies and legal notices with full AI reasoning chain and evidence snippets
- Designed per-decision compliance-oriented audit trail capturing confidence score, routing outcome, and AI reasoning for every triage decision

Nexus v4.3 — Agentic PKM Pipeline

[Personal Project]

Gemini API, Ollama (Qwen3-14B), MLX, FastAPI, MCP Server, LightRAG, DigitalOcean

- Designed two-stage Relay-Crystallizer architecture: cloud extraction (Gemini, DigitalOcean relay) → local Apple Silicon synthesis (Ollama/MLX) at \$0 local inference cost
- Built ingestion pipeline for YouTube, Web, PDF, Instagram, and clipboard inputs; autonomous multi-source research via Jina Search API with parallel `asyncio` processing across 4 simultaneous sources
- Deployed MCP server exposing semantic search and multi-hop graph retrieval (LightRAG) to Claude/Gemini CLI; enforced output reliability via strict Pydantic/JSON-schema validation, a 5-step JSON repair pipeline for local-model quirks, and prompt-level hallucination guarding (explicit source-content vs. training-knowledge separation)

“Accurate but Unreliable?” — Medical LLM Reliability Study

[MSc Team Paper, NCI, Apr 2026]

Fleiss' κ , Krippendorff's α , Mann-Whitney U

- Evaluated 3 medical LLMs (Aloe-Beta-8B, Meditron3-7B, Gemini-2.5-Flash-Lite) on 100 stratified USMLE-style questions $\times$ 10 runs $\times$ 2 languages = 6,000 total inferences
- Measured run-to-run consistency using Fleiss' κ , Krippendorff's α , bootstrapped 95% CIs, and Bonferroni-corrected Mann-Whitney U tests
- Identified a 23.3 κ -point Multilingual Reliability Gap in 8B specialists (EN → Hindi); relevant to reliability evaluation for high-risk multilingual clinical AI systems

ADDITIONAL ENGINEERING & ML WORK

MindScan (<i>MSc Team Paper — Co-authored, NCI, Apr 2026</i>): Co-developed parallel inference architecture training 15 and deploying 12 classifiers (XLM-RoBERTa, SVM, XGBoost) for depression classification and suicide risk detection; SVM F1=0.9269 on six-class type; XLM-RoBERTa F1=0.9993 on binary detection; deployed via Flask/Hugging Face Spaces
MindTune-OS (<i>MSc Module — Foundations of AI, NCI, Mar 2026</i>): Real-time EEG pipeline with online SGDClassifier (92.94% desktop accuracy); Groq/Llama3 agent selects Spotify interventions based on classified brain state; compressed to TinyML for edge inference on Arduino Uno R4 (74% accuracy, 32KB SRAM)

WORK EXPERIENCE

Oodlestack (Startup) <i>Lead Engineer</i>	Sep 2024 – Jan 2026 <i>India / Remote</i>
- Architected and delivered full-stack MVP for a B2B demand aggregation marketplace as sole engineer; owned product, backend, database, and frontend from architecture to prototype delivery - Built Python/FastAPI optimization service using Google OR-Tools (CP-SAT/MIP) to solve min-cost flow order allocation under MOQ, stock, pricing-tier, preferred-seller, and trust-score constraints - Designed Supabase (PostgreSQL) schema with SQL migrations, RPC functions, and triggers for trust scoring, association attribution, and recurring order dispatch	
HCL Technologies <i>Senior Software Engineer (SDET/QA Automation)</i>	Oct 2023 – Aug 2024 <i>Remote (India / Canada / EU)</i>
- Architected automated validation pipelines for Bank of Montreal’s capital markets platform; reduced regression cycles by 40% and accelerated releases by 2 weeks - Expanded coverage across 15+ critical trading workflows using Selenium/Java/TestNG; caught 23% more pre-production defects; shipped quality gates into Docker/GitHub CI/CD pipelines	
HCL Technologies <i>Software Engineer (QA Automation)</i>	Nov 2020 – Sep 2023 <i>India</i>
- Built automated validation frameworks for trading workflows; reduced manual testing by 60% with 85% automation coverage across 200+ test scripts - Contributed to Angular UI development; integrated test pipelines into CI/CD for trade creation and settlement workflows	

EDUCATION

National College of Ireland <i>MSc Artificial Intelligence — Academic Merit Scholarship</i>	Jan 2026 – Jan 2027 (Expected) <i>Dublin, Ireland</i>
SASTRA Deemed University <i>BCA Computer Applications</i>	Jan 2021 – Aug 2024 <i>India</i>
CGPA: 8.64/10 — First Class Distinction (Work-Integrated Learning)	
HCL TechBee, Shiv Nadar University <i>Full-Stack Web Development</i>	Jun 2019 – Nov 2020 <i>India</i>

CERTIFICATIONS & LANGUAGES

Certifications: Capital Markets Analyst (CMSA) – CFI (Oct 2023) Agile Project Management – Google (Jan 2023) Data Science Foundations – IBM (Dec 2020) NASA Open Science 101 (Sep 2025)
Languages: English: C1 (IELTS 7.0) Malayalam: Native Hindi/Tamil: C1