

Mohit Kumar Dubey

Head of AI, ML & Engineering

AI Strategy • Generative & Agentic AI • LLM Platforms • Engineering Leadership • Enterprise AI

Kuala Lumpur, Malaysia • +91 9611886748 • mohitdubey60@gmail.com

linkedin.com/in/mohit-kumar-dubey • huggingface.co/EntermindAI

Executive Summary

AI and engineering leader with 16+ years building AI-first products and the organizations that ship them. Currently Head of AI, ML & Engineering at a Kuala Lumpur AI consultancy serving enterprise and government clients across SEA and MENA. Previously led a 35+ engineer organization at VerSe Innovation (500+ employees), India's largest short-video and news platform at 350M+ users. Combines production GenAI delivery with org design, Engg P&L ownership (licenses, new hires, appraisals), AI governance and executive stakeholder management across the Gulf, SEA and the US.

Selected Impact

3x delivery output at flat headcount: LLM automation of recurring workflows company-wide (Entermind)
25% loan disbursement conversion uplift, 20% retention gain: BuddySmart conversational AI and agentic RAG (VerSe)
23% reduction in lending default rates: credit-risk scoring models for lending partners (VerSe)
30 to 40% cut in LLM API cost: model routing and caching, DailyTatva RAG v2 (VerSe)
30% crash-rate reduction, 65% faster fix turnaround: anomaly-detection pipelines at 350M+ users (VerSe)
45% lift in AI brand citation scores: Generative Engine Optimization platform for GXS, Grab Group (Entermind)
35+ engineers led across four disciplines: AI/ML, backend, mobile and web within a 500+ employee company

Core Competencies

Leadership: AI Strategy & Transformation • Engineering Organization Design • Technology Portfolio Management • Cross-Functional Leadership • Talent Development

Business: Engg P&L Ownership • AI Product Strategy • Enterprise Sales Support • Go-To-Market Enablement • Strategic Partnerships

Governance: AI Governance • Responsible AI • AI Operating Models • Enterprise AI Adoption • Stakeholder Management

Technical Skills

AI & ML: LLMs, RAG, Agentic RAG, Multi-Agent Systems, Fine-tuning (LoRA, PEFT), NLP, Recommendation Systems, AI Harness, Evaluation Harness, Agentic Loop design, Prompt Engineering, MLOps, Credit-Risk Models, Structured Outputs

Agent Frameworks: LangChain, LangGraph, CrewAI, DeepAgents

Training & Serving: PyTorch, TensorFlow, vLLM, SGLang, FastAPI

Model Platforms: AWS Bedrock, Amazon SageMaker, Vertex AI, Azure AI Foundry, OpenAI, Gemini, Claude, Cohere

Voice & Multimodal: ElevenLabs, Tavus, LiveKit, Pipecat, Daily, Serper, Tavily

Data & Vector Stores: PostgreSQL, pgvector, Qdrant, MongoDB, Chroma, FAISS, Redis

MLOps & Observability: Langfuse, Phoenix, Opik, LiteLLM, MLflow, Kubeflow, Airflow

Infrastructure: AWS, GCP, Docker, Kubernetes, ECS, EC2, CI/CD

Languages: Python, Pandas, NumPy, Pydantic, SQL, Swift

Professional Experience

Entermind Malaysia Sdn. Bhd., Kuala Lumpur, Malaysia

Jan 2026 - Present

Head of AI, ML & Engineering

AI consultancy serving enterprise and government clients across SEA and MENA. Lead an 11-engineer team across backend, AI/ML, frontend and QA, and own the company-wide AI and engineering roadmap.

- Drove a **3x increase in delivery output** with flat headcount by identifying recurring workflows across the business and automating them with LLM pipelines.
- Cut the Chief Credit Officer's routine meeting load **20%** and improved information availability **40%** at **First Abu Dhabi Bank** by architecting an AI Digital Twin with enterprise knowledge, voice, memory and agentic workflows.
- Lifted **GXS (Grab Group)** AI brand citation scores **45%** across ChatGPT, Gemini, Claude and Perplexity by building a Generative Engine Optimization platform.
- Improved pre-sales effectiveness **30%** and cut claims and HR validation effort **45%** by deploying enterprise RAG and conversational AI; automated leave management.
- Launched **Rukun Ready**, a fine-tuned 32B guardrail LLM (Qwen2.5-32B, LoRA/PEFT) aligned to Malaysia's Rukun Negara principles; the firm's first open-source AI asset, now at **3K+ HuggingFace downloads**.
- Lead **11 engineers** across backend, AI/ML, frontend and QA, owning hiring, capability planning and delivery allocation across concurrent enterprise engagements.
- Established the enterprise AI operating model: technical review gates, architecture sign-off and delivery prioritization across concurrent engagements.
- Built a reusable GenAI capability layer (RAG ingestion, evaluation harness, guardrail scoring) so client engagements begin

from proven assets rather than from scratch.

- Primary technical point of contact for enterprise clients across MENA and SEA, from pre-sales discovery to delivery sign-off.

VerSe Innovation, Bengaluru, India

Jun 2025 - Nov 2025

Director of Engineering

*Dailyhunt/ Josh/ BuddyLoan - Continued as de facto Head of Engineering for BuddyLoan on promotion to Director, **scaling the org from 20+ to 35+ engineers** within a 500+ employee company*

- Lifted loan disbursal conversion **25%** and retention **20%** by shipping **BuddySmart**, a production Conversational AI and Agentic RAG platform for personalized loan matchmaking with real-time eligibility checks.
- Improved onboarding conversion **35%** and cut drop-off **20%** by redesigning acquisition funnels and rebuilding end-to-end telemetry.
- Deployed AI-driven financial insights for lending partners and offer personalization (PA/PQ) journeys, extending the platform's monetization surface.
- Stabilized and scaled the BuddyLoan app and backend platform, aligning mobile, web, backend and AI/ML roadmaps behind a single delivery plan.
- Directed ML model design for customer credit-risk scoring, connecting model performance to lending economics reviewed with business leadership.
- Directed **35+ engineers** across AI/ML, backend, mobile and web; restructured ownership and instituted operating reviews that improved delivery cadence.

VerSe Innovation, Bengaluru, India

Nov 2023 - May 2025

Associate Director of Engineering, AI & Platform

*Dailyhunt/ Josh/ BuddyLoan - Took on de facto Head of Engineering for the BuddyLoan lending vertical, **leading 20+ engineers**, while owning the AI/ML roadmap across content and lending for a 350M+ user platform.*

- Reduced LLM API cost **30 to 40%** while improving retrieval quality by delivering **DailyTatva RAG v2** with model routing and response caching.
- Cut lending default rates **23%** and improved model accuracy **15%** by deploying credit-risk scoring models for lending partners.
- Increased search-news engagement **15%** and cut manual editorial curation by launching a multi-agent news summarization system on LangGraph.
- Owned intent identification, prompt optimization and LLM cost governance across all GenAI products, establishing per-feature unit economics.
- Shipped **14 user-acquisition services** plus third-party integrations and a custom ads framework, and ran offshore client engagement for the platform.
- Sustained **99.8% crash-free sessions**, 85%+ notification-to-play and 90%+ launch-to-play across Josh and Dailyhunt.

VerSe Innovation, Bengaluru, India

Apr 2021 - Oct 2023

Software Architect, AI & Platform Architecture

Dailyhunt/ Josh

- Reduced crash rate **30%** and crash-fix turnaround **65%** by designing anomaly-detection and crash-analysis pipelines across 350M+ user platforms.
- Delivered **DailyTatva RAG v1**, the retrieval and inference foundation for every subsequent GenAI news product.
- Built a GenAI-powered application evaluator for crash prevention and a decisioning service validating CTR/CTA against third-party mappings.
- Drove cross-platform evaluation (React Native, Kotlin Multiplatform) to inform long-run mobile architecture strategy.

VerSe Innovation, Bengaluru, India

Apr 2018 - Mar 2021

Principal Software Engineer, ML & Recommendations

Dailyhunt/ Josh

- Improved multilingual personalization across **14+ Indian languages** for 350M+ users by building production ML for audio language detection and video similarity scoring.
- Engineered modular SDKs for networking, ads and video streaming (custom FMP4 and M3U8 with caching), establishing reusable architecture across both apps.
- Integrated hooks for ML-based A/B testing and content recommendations, shortening integration time for later inference services.

VerSe Innovation, Bengaluru, India

Jun 2016 - Mar 2018

Lead Software Engineer

Dailyhunt

- Achieved **99.8% crash-free sessions**, **50% video playback improvement** and **68% push-analytics accuracy gain** by leading iOS modernization from Objective-C to Swift with MVVM-C.
- Led React Native (Android and iOS) delivery and Node.js acquisition services alongside Windows 10 UWP development, then transitioned into AI/ML delivering recommendation and language detection.

Torry Harris Business Solutions, Bengaluru, India

Apr 2014 - May 2016

Senior Software Engineer, Enterprise & Government Mobility

Enterprise and government mobility delivery across the GCC: Dubai Airports, Bahrain Government, Kuwait ministries.

- Delivered **14+ enterprise and government platforms** and led the entire **BU4** across multiple teams, pre-sales, technical design reviews and CMMI compliance.
- Developed services for Bahrain Government, Kuwait Ministries and Dubai Airports.

GlobalLogic, Nagpur, India
Senior Software Engineer

Nov 2012 - Apr 2014

- Delivered SmartTV, Xbox and Windows 8 platform applications, including a Verizon movie-rental kiosk platform.
- Developed a banking-grade OTP encryption SDK - Gemalto Singapore (onsite).

Levitte Mobile Technologies, Gurugram, India
Software Engineer

May 2012 - Oct 2012

Delivered consumer-facing applications for Jubilant Food and Religare Wellness with analytics-driven engagement and personalization.

Ritusha Consultants Pvt. Ltd., Kanpur, India
Corporate Trainer

May 2009 - Apr 2011

Trained **100+ engineers** across TCS fresher batches and IIIT-G cohorts while building custom service-based software for enterprise clients.

Selected AI Projects

Mindara, Entermind's AI Employee for Enterprise Engagement. Built an always-on conversational AI colleague that replaces the form with adaptive follow-up questioning, surfacing real requirements and handing structured briefs to sales: **15% increase in solutioning lead calls**. Architected across seven capability levels, Level 1 in production, voice built and staged for release.

Stack: LiteLLM gateway, Claude Opus, OpenAI, Gemini, Langfuse, ElevenLabs, LangGraph

Rukun Ready, Guardrail LLM for Policy Validation. Enterprise clients needed LLM outputs checked against national policy principles before deployment. Fine-tuned Qwen2.5-32B to score outputs and emit structured JSON compliance verdicts with safe rewrites; shipped as the firm's first open-source AI asset, now at **3K+ HuggingFace downloads** and used as a pre-sales credibility anchor.

Stack: Qwen2.5-32B, LoRA, PEFT, vLLM, RunPod, HuggingFace

AI Digital Twin, First Abu Dhabi Bank. The Chief Credit Officer's calendar was the bottleneck for routine credit queries. Built a multimodal conversational twin over enterprise knowledge with voice, memory and agentic workflows, giving stakeholders 24/7 access: **20% lower routine meeting load** and **40% better information availability**.

Stack: Azure AI Foundry, LangGraph, Tavus, ElevenLabs, Pipecat, Daily, pgvector, Redis, Azure AI Search, PostgreSQL

GEO Platform, GXS (Grab Group). Buyers had shifted discovery to AI assistants where the brand was largely invisible. Built citation-aware analytics and optimization across four assistants with competitor benchmarking, lifting **AI brand citation scores 45%** on ChatGPT, Gemini, Claude and Perplexity.

Stack: AWS Bedrock, OpenAI, Gemini, Claude and Perplexity APIs, LangGraph, PostgreSQL, EC2, Docker

BuddySmart, Conversational AI and Agentic RAG. Loan discovery was form-driven with heavy drop-off. Built a conversational matchmaking assistant with real-time eligibility checks and personalized offer delivery, which became the platform's primary conversion lever: **25% disbursal conversion uplift** and **20% retention gain**.

Stack: LangGraph, LangChain, Cohere Embeddings, Qdrant, AWS Bedrock, ECS, Docker

DailyTatva RAG v1 and v2. Personalized news retrieval at 350M+ user scale with LLM spend growing faster than usage. Built the v1 retrieval and inference foundation, then added model routing and caching in v2 to cut **LLM API cost 30 to 40%** while improving retrieval quality.

Stack: LangGraph, LangChain, Gemini, Cohere Embeddings, Vertex AI, Qdrant, ECS, Docker

Multimodal Enterprise Knowledge Base. Sales and delivery teams could not retrieve across years of decks and internal documents. Built a multimodal RAG pipeline ingesting PPT, PDF and Word with semantic search at scale, contributing to a **30% pre-sales effectiveness gain**.

Stack: Gemini 2.5 Multimodal, OCR, pgvector, PostgreSQL, Vertex AI, ECS, Docker

Education

Executive Post Graduate Programme in Machine Learning & AI, Specialization: Generative AI
 International Institute of Information Technology (IIIT), Bangalore, India

Nov 2024

Bachelor of Science, Information Technology
 Kanpur University, India

Mar 2009

Certifications: AWS Certified AI Practitioner • Responsible AI and Advanced Generative AI (DeepLearning.AI) • Mastering System Design • Software Architecture: Patterns for Developers • SOLID Principles in Software Design • REST-based Microservices API Development in Golang

Languages: English, Hindi • **Exposure:** Malaysia, Singapore, UAE, India • **Awards:** Integrated Team Award, Team Excellence (x2), Reward & Recognition (VerSe)