

Uday Dahiya

+91-93197 04550 | udaydahiya4601@gmail.com | [linkedin.com/in/uday-dahiya](https://www.linkedin.com/in/uday-dahiya) | udaydahiya.com

EXPERIENCE

nStore Retech Private Limited

Jun 2024 – Present

B2B SaaS

Mumbai, India

AI Solutions Architect

- Owned product strategy & roadmap for AI products deployed by banks & enterprises to their customers; running discovery & prioritization to translate user pain points & **RBI FREE-AI Framework** tradeoffs into a shipped roadmap
- Designed the product & system architecture for an **agentic-AI application in personal finance**, spanning a deterministic metrics engine, scoring, nudge & conversational agent layers, & led development & deployment end to end
- Led **client-facing solutioning** for a mid-size bank across Wealth Management, Credit Risk, Cross-Sell & Churn, pairing financial & AI lenses from discovery to solution design to map value levers worth **INR 100 Cr+** of P&L impact
- Led HLD & LLD reviews with engineering to architect scalable **RAG, agent memory & tool-calling** into KYC & credit-risk operations pipelines for bank clients, shipping to production & improving operational efficiency by **30%**
- Bridged the CXO suite & engineering teams to drive internal AI adoption, cutting delivery & support TAT by **60%**
- Designed & shipped the company's **internal AI system**, integrated with Jira, GitHub & M365 for in-workflow adoption

Customer Experience Manager

- Owned customer experience & product metrics end to end for 6 enterprise clients serving **1Cr+ end customers** on a B2B2C white-label platform, building the **measurement framework from zero** across the full customer funnel
- Ran discovery with enterprise client leadership to define & track activation, engagement, retention, cross-sell / upsell & CSAT, translating pain points into prioritized requirements & **50+ features shipped** with product & engineering
- Designed & shipped an **AI-based customer-resolution system** end to end, cutting resolution time by **75%** & automating support & operations workflows for clients' end customers across all 6 enterprise accounts

PROJECTS AND HACKATHONS

Regional Ad Factory | *FastAPI, BytePlus ModelArk, Sarvam AI, Seedance, ffmpeg*

Aug 2026

- Built an **AI video-ad factory** turning one product photo into a ready-to-run ad voiced natively in **Indian regional languages**, letting e-commerce & consumer brands localize campaigns per market at **under \$1 per ad**
- Designed a **plugin architecture** with four adapter seams (LLM, audio, video, use-case) resolved via a registry, so a new language, ad format or AI vendor is a config change, not a rebuild, backed by **395 tests** that run with zero network

alphaXiv x marimo | *AI Research Paper Hackathon*

July 2026

- Implemented the **Hierarchical Reasoning Model** as an interactive notebook, a **27M-parameter** net splitting thinking into a slow planner & fast executor to solve puzzles from 1,000 examples, showing reasoning comes from design, not scale
- Trained the **Tiny Recursive Model**, one **7M-parameter** network looping over & revising its own answer to out-score billion-parameter LLMs, pointing to reasoning that runs on modest on-prem hardware, not a frontier budget

MCP Gateway | *Python, FastMCP, Model Context Protocol*

Oct 2025

- Consolidated multiple MCP servers behind one governed endpoint with an **enterprise MCP gateway**, auto-discovering tools at runtime & adding hook-based plugins for guardrails, tracing, PII detection & pre-load security scans of each server

AWS SuperHack | *Agentic AI for IT Operations Management (Google RLM)*

Sept 2025 – Oct 2025

- Led cross-functional team in developing & deploying **AI alert triage system** using Regression Language Models (RLM/T5) for text-to-text regression on raw IT data, achieving **90%** noise suppression for Managed Service Providers
- Deflected **40%** of support tickets with a production **RAG chatbot** built on an open-weight 7B model (Qwen 2.5) & ChromaDB, indexing 30+ articles so users self-resolve via semantic search before a ticket reaches an engineer

American Express | *ML Super Bowl Challenge*

July 2023 – Oct 2023

- Ranked **Top 10 of 5,500 teams** with a **merchant recommendation engine** built on XGBoost over a 12M+ row commercial dataset, one-hot encoding 500+ categorical features, compressing via PCA & tuning 145 trees with 3-fold CV

EDUCATION

Indian Institute of Management, Indore

Indore, India

Master of Business Administration (MBA)

Mar 2024

BA, Foundations in Management

Apr 2022

TECHNICAL SKILLS

Product Management: PRDs, Roadmaps, Strategy & Tradeoffs, User Journeys, Prioritization, Agile, Scrum & Jira

ML & AI: RAG, Vector Embeddings, SLM/LLM, Agentic AI & MCP, Prompt Engineering, Claude Code & Replit

Full-Stack Development: React, Node.js, Python, SQL, FastAPI, APIs, Docker, AWS & Azure, HLD & LLD Design

AWARDS AND CERTIFICATIONS

- Microsoft Certified: Azure AI Fundamentals (AI-900)** - Azure AI workloads, computer vision, NLP & responsible AI
- Anthropic Claude Developer** - Building with the Claude API, Claude Code in Action, Claude on Google Cloud & Bedrock
- Anthropic Agents & MCP** - Model Context Protocol (Intro & Advanced), Agent Skills, AI Fluency Framework
- Awarded by **AmEx** as the national finalists out of 5500 teams in the Analytics Campus Superbowl Challenge