SHYAM PADIA

+1 (857)-294-9264 — shyam.padia930@gmail.com — Boston, MA
crypticsoul.dev — github.com/crizzy9 — linkedin.com/in/shyam-padia

Summary

AI technical lead building agent systems, retrieval and recommendation platforms, and large-scale data products. Combines hands-on architecture with engineering leadership to turn model capabilities into practical workflows, improve reliability, and reduce operating cost. Work spans conversational planning, marketing automation, retrieval, personalization, distributed data processing, and full-stack products. Independent engineering includes declarative infrastructure, mobile applications, robotics, language models, and owner-controlled knowledge systems.

Education

Northeastern University, Khoury College of Computer Sciences — *Boston, MA* — **Sept 2017 - Dec 2019**
Master of Science in Computer Science — GPA: 3.2/4
University of Mumbai, Fr. Conceicao Rodrigues College of Engineering — *Mumbai, India* — **Aug 2012 - June 2016**
Bachelor of Engineering in Computer Engineering — GPA: 7.4/10
Coursework: Natural language processing, Reinforcement learning, Robotics, Machine learning, Artificial intelligence

Work Experience

MosaicApp — AI Technical Lead — *San Francisco, CA* — **Aug 2026 - Present**
- Rearchitected Chello's AI execution layer, improving performance by 25% while reducing costs by 60%.
- Expanded Chello's Model Context Protocol (MCP) capabilities across connected systems, delivering 27 tools for agents to retrieve and update planning data through a shared interface.
- Redesigned agent caching to reuse prior work and reduce repeated processing within the LangChain harness.
- Extended the LangChain harness with subagents and task management to coordinate multi-step planning workflows.
- Delivered Slack and Microsoft Teams integrations, bringing Chello's planning capabilities into team collaboration workflows.
- Introduced query-complexity-based model routing to reduce token usage and inference cost.
- Enabled tiered billing for Chello by integrating Stripe and MonetizeNow.

Intuit, Personalization — Senior Software Engineer — *Mountain View, CA* — **Jan 2025 - Aug 2026**
- Built a LangChain agent orchestrator for RTB, on-call, and DevOps workflows.
- Architected unified user profiles with multi-source ingestion and real-time user embeddings on GPU-optimized EMR.
- Led a cross-regional team building a unified recommendation platform with Vespa, gRPC, and OpenSearch, replacing legacy systems.
- Developed an AI marketing agent with a six-person team using SymphonyAI, LangChain and MCP, cutting campaign go-live time by 90%.
- Led five engineers building a bulk content system with dynamic templates and audience segmentation, increasing capacity by 1,400%.
- Built an LLM-powered content pipeline that automated marketing workflows and improved time-to-market by 75%.
- Created an MCP server for AI-powered marketing campaign creation, real-time content generation, and automated deployment.
- Defined and presented the personalization AI strategy across content generation, recommendations, predictive analytics, and workflow automation.

Intuit, Personalization — Software Engineer 2 — *Mountain View, CA* — **Jan 2021 - Jan 2025**
- Engineered a self-service analytics platform with PySpark, AWS EMR and React, reducing incidents and impression loss by 40% and increasing marketer productivity by 50%.
- Developed a RAG assistant with Hugging Face and LangChain for marketing content generation, reducing authoring time.
- Designed scalable PySpark data entities for real-time multi-source ingestion, eliminating approval bottlenecks.
- Implemented AWS Personalize on clickstream data with 20% lower API latency.
- Built a Kafka event pipeline for targeted offers that generated $10M in incremental revenue through dynamic audience targeting.
- Trained and deployed XGBoost and neural-network models for predictive business-health scoring, reducing investment risk by 23%.

Intuit, QuickBooks — Software Engineer 1 — *Mountain View, CA* — **Jan 2020 - Jan 2021**
- Built a predictive revenue-forecasting model using customer behavior to estimate the impact of dynamic pricing strategies.
- Engineered a segmentation pipeline with AWS Glue and PyArrow to identify eligible cohorts for targeted pricing campaigns.
- Integrated price simulation and campaign execution components, delivering $20M in cost savings.
- Developed a pricing management platform with React, Spring Boot, and AWS, increasing operational efficiency by 50%.
- Implemented a parallelized synthetic-testing framework with Cypress, preventing $10M in potential revenue loss.

Intuit, TurboTax — Software Engineer Intern — *San Diego, CA* — **May 2019 - Aug 2019**
- Developed OCR document processing for W2 uploads in TurboTax onboarding, reducing tax preparation time by 20%.
- Integrated an NLP intent-classification decision engine, replacing a third-party API and reducing latency by 30%.

Intuit, TurboTax Software Engineer Co-op *San Diego, CA* **May 2018 - Dec 2018**

- Engineered a rule-based case-deduplication service for the Intuit Help Portal, improving response time by 45% and reducing duplicates by 7%.
- Developed Smart Fields with React and PHP to automate content publishing, reducing publishing time by 30 minutes and support cases by 10%.
- Architected a serverless URL-redirect service with AWS Lambda and DynamoDB, improving routing efficiency by 35%.

OutreachCircle Software Developer *Bangalore, India* **Nov 2016 - May 2017**

- Integrated Elasticsearch contact matching for voter registration and social profiles, improving outreach by 10%.
- Developed a campaign outreach platform with Meteor.js and Angular, improving canvassing efficiency and voter turnout by 15%.

Technical Skills

AI engineering:	LangChain, MCP, Hugging Face, Vespa, RAG, Model routing, Agent orchestration, AWS Personalize
Machine learning & vision:	PyTorch, TensorFlow, OpenCV, scikit-learn, XGBoost, Keras, NLTK, Gensim, Reinforcement learning
Languages:	Python, TypeScript, Java, Go, JavaScript, SQL, Scala, C / C++, Nix, Lua
Web & service development:	React, Spring Boot, HTMX, gRPC, Astro, Django, Node.js, REST, Cypress
Data engineering:	PySpark, Kafka, AWS EMR, PyArrow, Hadoop, Pandas, Hive, Beam, Flink, EventBus, RabbitMQ
Cloud & infrastructure:	AWS, Docker, Kubernetes, NixOS, SageMaker, Lambda, Glue, Athena, Redshift, Nix Flakes, Home Manager
Storage & search:	PostgreSQL, DynamoDB, Redis, OpenSearch, MySQL, SQLite, MongoDB, Elasticsearch, Vector databases

Projects

Lumino **Declarative infrastructure**

- Developed a declarative NixOS platform for desktops and self-hosted services, keeping infrastructure and service configuration reproducible and recoverable.
- Unified host, service and user setup through YAML-driven modules, using Nix Flakes, Home Manager and sops-nix with version-controlled configuration.

Naavik **AI-powered career workflows**

- Built a self-hosted career platform with profile management, explainable job matching, and tailored resume and cover-letter generation using Python, FastAPI, HTMX, PostgreSQL and Typst.
- Separated reviewed preparation and dry runs from confirmed application submission, recording submission receipts and uncertain outcomes to support deliberate follow-up.
- Extending candidate outreach, interview tracking and application workflows while V1 release qualification remains ongoing.

Plano **Flight tracking — In development**

- Built a Go/SQLite backend paired with an Expo/React Native mobile app for flight status, inbound aircraft, telemetry and travel history.
- Provided a credential-free synthetic demo; live-source and device qualification remain in progress.

Omnidirectional Autonomous Robot **Robotics & computer vision**

- Engineered a Raspberry Pi robot with a servo-controlled 6-DOF arm, inverse kinematics, OpenCV object detection, speech-to-text commands, autonomous navigation, and obstacle avoidance.

Reinforcement Learning for Robotic Manipulation **Transfer-learning research**

- Compared Q-Learning, DQN, Policy Gradient, and MCMC for robot-arm simulation using PyTorch and OpenAI Gym. Transfer-learning experiments achieved 35% faster convergence and 28% improved task performance.

Multi-Track Music Generation **Transformer architectures**

- Developed separate transformer models for melody, bass, drums, and harmony, connected by custom global multi-headed attention. The polyphonic generation work improved musical coherence by 45%.

Paragraph Embeddings with Variational Autoencoders **Natural language processing**

- Implemented TensorFlow architectures for unsupervised paragraph representations on the Gutenberg Corpus, comparing VAE and LSTM Seq2Seq approaches and improving semantic-similarity performance by 22%.

Fan Forge — Comic-Con Museum **Community software & team leadership**

- Led a backend team of five building an open-source application where fans could curate exhibit and artifact ideas for the Comic-Con Museum in San Diego, using Express.js and AWS Lambda.

Avid **Development preview**

- Built a personal reading and audiobook application using React and TypeScript, including EPUB, PDF and comic readers, BookOrbit integration, and local EPUB-to-audiobook alignment.
- Uses Capacitor mobile and Tauri desktop hosts; device qualification and broader offline support remain in progress.

PupCam **Research**

- Researching an owner-controlled pet camera with self-hosted media and intelligence.
- Evaluating hardware feasibility, access and recovery constraints, and a practical feature scope before implementation.

Search Engine **Information retrieval**

- Implemented a multiprocessing search engine, web crawler, and snippet generator combining BM25, smoothed query likelihood and TF-IDF retrieval models.

Automatic Hypernym Detection **Natural language processing**

- Trained matrix-factorization models on the Gutenberg corpus to extract hypernyms from linguistic patterns using NLTK and Gensim.

StarCraft Deep Learning Bot **Computer vision and game agents**

- Developed a TensorFlow CNN and OpenCV-based StarCraft bot to process game frames, recognize units and resources, and support map awareness and unit control.

ROS2 Simulation Training **Robotics experimentation**

- Experimented with ROS2 simulation tools for robotics.

Obsidian Knowledge System **Personal knowledge engineering**

- Created an AI-assisted Obsidian knowledge workflow using open-source plugins to organize notes and support connected retrieval.

Home Assistant with Ollama **Local AI and automation**

- Explored local Ollama integration with Home Assistant and Matter-based automation workflows.

Multi-Agent Workflow Integration **Agent orchestration**

- Configured a Hermes-based multi-agent workflow for task coordination and home-automation integration.

Corne ZMK Keyboard **Hardware and firmware**

- Built and soldered a Corne keyboard with custom ZMK firmware, nice!nano controllers and nice!view displays.

Certifications

Machine Learning Professional Certificate - General Assembly **June 2022**

- Supervised and unsupervised learning, neural networks, ensembles, model evaluation, and deployment strategies through hands-on Python work.

Apache PySpark Certification - Databricks **Dec 2021**

- Distributed processing, Spark SQL, DataFrames, performance optimization, and production-scale data pipelines on Databricks.

AWS Certified Solutions Architect - Associate - Amazon Web Services **Oct 2019**

- Distributed systems, cloud architecture, service selection, cost optimization, and scalable applications on AWS.

Open Source Contributions

Vicinae **2025**

- Contributed features to the C++ and Qt desktop launcher, including Raycast-compatible extension APIs, TypeScript SDK integration, and Linux-native keyboard-driven productivity tools.

Mopidy **2024**

- Contributed features and fixes for Spotify integration, playlist synchronization, and audio-stream optimization for Raspberry Pi deployments.