

ARCHIT AGRAWAL

San Francisco Bay Area | 623-312-0435 | architagrawal000@gmail.com | [LinkedIn](#) | [GitHub](#) | [Personal Website](#)

PROFESSIONAL EXPERIENCE

EdPlus

AI Software Engineer

05/2026 - Present

Phoenix, AZ

- Own the architecture, roadmap and delivery of a 12-agent AI analytics platform on AWS Bedrock and Step Functions that turns any structured dataset into cited, verified answers; TypeScript, NestJS, Nuxt.
- Cut wrong answers 85% on 164 published benchmark figures by reading every answer back in plain words and verifying it against the question before display.
- Extended analysis to statistical and causal questions with 25 tests and six causal estimators validated against scipy and statsmodels, scoring 68% on StatQA versus GPT-4o's best reported 64.83%.
- Quadrupled the insights automated reports surface, 13% to 52–59% on unseen datasets, by having an LLM agent choose the analyses while deterministic code computes every number.
- Raised accuracy on the hardest benchmark questions from 60% to 93% with zero wrong answers, by routing every component through one shared semantic layer after larger models did not help.
- Eliminated a silent failure mislabeling 1 in 8 rows in the prior pipeline, reaching 100% join integrity with schema-constrained LLM output.
- Scaled labeling as a distributed Step Functions fan-out across 64 parallel lanes with a 5% failure circuit breaker, after tracing silent row loss in a run that reported success.
- Showed model choice moves accuracy 33x more than pipeline design across models costing \$5–\$89 per 100k rows, redirecting the team's recommendation from accuracy to reliability.
- Benchmarked the platform against Claude Opus 5.5 working blind on public survey data, cutting wrong charts 93% with judge-free rule checks and an expert-calibrated LLM judge.
- Built a plain-English chart experience in Nuxt where readers stack, sort, filter, take top-N and undo directly, backed by in-process DuckDB that held latency flat as rows grew 200x.
- Designed a chart recommender over 17 chart types grounded in visualization research, accepting an LLM suggestion only when it outscores the rules.
- Sped up a full AWS pipeline run 6x, 500 to 85 seconds, by replacing managed agent-runtime stages with direct tool calls once benchmarks showed each made one call and stopped.
- Kept deterministic rules wherever an agent failed to beat them: no LLM chart proposal outsourced the rule table, and two managed agent runtimes were removed after re-benchmarking.
- Changed the evaluation from scoring charts to scoring the questions a survey asked, after healthy chart scores hid that only 19% were answered; coverage rose to 78%.

MyStage Music Inc.

Founding AI/ML Engineer

07/2025 - 06/2026

Remote

- Re-architected a 14-Cloud-Function backend into one checkpointed LangGraph agent in Python on GCP Cloud Run, cutting production service dependencies 64%.
- Scaled ingestion to 70,000+ records a day from 650+ venues with a Playwright pipeline, transactional locking and Gemini entity resolution that raised search-dataset accuracy 25%.
- Delivered sub-50 ms search on FastAPI and Algolia with real-time Firestore sync, 200+ pytest tests and Logfire tracing.
- Implemented pause-and-resume on LangGraph interrupts, so a record missing context waits at its checkpoint instead of failing the batch.
- Guaranteed retry safety with deterministic task IDs and create-if-absent writes, so a retried job can never overwrite one already in flight.
- Designed a reusable scraping subgraph that parent graphs compose without checkpoint conflicts, letting webhook-triggered scrapes skip the fetch step.

EdPlus

Software Engineer

09/2023 - 05/2025

Phoenix, AZ

- Led a multi-tenant RAG assistant, React to retrieval layer, used by 1,000+ faculty to build courses for 60,000+ students.
- Cut transcript analysis 16x, four hours to 15 minutes, with a Neo4j knowledge graph queried by validated LLM-written Cypher.
- Isolated each college's data at the retrieval layer, so no tenant's material can surface in another's and onboarding became a config change, not a deployment.
- Built a Prompt Flow eval harness scoring groundedness, relevance and coherence to gate every prompt change on results.
- Added hybrid keyword and vector search so exact course codes resolve correctly where embeddings alone blur them together.
- Attached a source citation to every answer, linking document and section, so faculty can verify a claim instead of trusting it.

Arizona State University*Student Researcher***08/2024 - 05/2025***Phoenix, AZ*

- Built a PyTorch and Diffusers benchmark for text-to-video models that scores physical plausibility separately from visual quality, holding seed, resolution and sampler fixed.
- Found visual quality barely predicts physical plausibility: the top model by human preference ranked last on causal consistency.
- Designed a frame-level rubric for rigid-body motion, gravity, momentum and occlusion, and measured inter-rater agreement before trusting any aggregate score.

Knowledge Exchange for Resilience, Arizona State University*Data Research Aide***06/2024 - 08/2024***Phoenix, AZ*

- Cut p95 retrieval latency 60% at 95% recall with a hybrid FAISS and Neo4j search layer.
- Built a FastAPI and pgvector recommender reaching 85% top-k precision for collaborator matching, fed by an ETL ingesting 10,000+ profiles a day, plus 20+ REST APIs on .NET 8.
- Lowered API latency 90% with Redis caching and SQL rewrites, and raised dataset accuracy to 95% with PostgreSQL constraints.

Zeus Learning*Software Engineer***01/2022 - 07/2023***Mumbai, India*

- Built a Redis-backed desk-demand forecasting service used in 300+ Fortune 500 offices, with reported occupancy up 30%.
- Cut deploy time 70% and production incidents 40% by containerizing the CI/CD release pipeline and adding SonarQube gates.
- Migrated a C# .NET monolith to Dockerized Kubernetes microservices, reducing infrastructure cost 20% and footprint 35%.

EAT.FIT*Product Intern***09/2021 - 12/2021***Bengaluru, India*

- Cut driver-tracking server cost 45% with distance-based polling and client-side interpolation that kept map movement smooth.

DA-IICT*Computer Vision Researcher***05/2021 - 08/2021***Gandhinagar, India*

- Engineered an autonomous-driving perception stack in Python: radar and lidar capture, timestamp alignment, extrinsic calibration and point-cloud processing, with a replay harness over recorded drives.

PROJECTS**SRP Electric MCP Server | [GitHub](#)****12/2025 - 01/2026**

- Built a Python MCP server on FastMCP that logs into an undocumented utility portal and gives AI agents 7 tools and 4 analysis prompts for billing, usage, payments and outages, with a one-click setup wizard.
- Reverse-engineered the portal's login flow and API calls, since the utility publishes no API documentation.

MCP GitHub PR Review Agent**07/2025 - 08/2025**

- Built a TypeScript MCP service that reviews pull requests from the diff, linked ticket and CI result, returning schema-typed inline findings that gate merges by category.
- Reviews large PRs file by file with a per-file verdict, because one 8,000-line diff in a single prompt produces a summary, not a review.

AiJockey - AI DJ Pipeline | [GitHub](#)**2025 - Present**

- Built an end-to-end AI DJ in PyTorch and FastAPI: stem separation, an LLM transition planner, 25+ DSP transitions and LUFS mastering, ported off CUDA to an AMD MI300X.
- Trained a MERT-95M reward head predicting four-axis audio aesthetics (final MSE 0.127) so the segment picker scores candidates without running full inference per render.

TECHNICAL SKILLS**Languages:** Python, TypeScript, JavaScript, SQL, C#**Frameworks:** FastAPI, NestJS, Node.js, .NET, React, Vue/Nuxt, Next.js, LangGraph, PyTorch**AI & Data:** RAG, LLM evaluation, MCP, AWS Bedrock, PostgreSQL, pgvector, FAISS, Neo4j, Redis, DuckDB, ETL pipelines**Cloud & DevOps:** AWS (Step Functions, Lambda, S3, CDK), GCP Cloud Run, Docker, Kubernetes, CI/CD**EDUCATION****Arizona State University (ASU)***Master of Science, Computer Science***08/2023 - 05/2025***Tempe, AZ, USA*

- **GPA:** 4.0

Dhirubhai Ambani Institute of Information and Communication Technology*Bachelor of Technology, Information and Communication Technology***08/2018 - 05/2022***India*