

ANKIT DHANDHARIA

AI Engineer

New Jersey, USA | (646) 630-6409 | Dhandhariaankit03@gmail.com | [Linkedin](#) | [Github](#)

PROFESSIONAL SUMMARY

AI Engineer building LLM applications and agents at JPMorgan Chase Business Banking with OpenAI and Anthropic Claude APIs, MCP, and RAG on AWS. 4+ years in software and ML engineering, including Azure OpenAI and LangChain, work for a bank client at InfoWay Solutions. Ships GenAI features with the evals, guardrails, and audit logs Model Risk review expects.

TECHNICAL SKILLS

LLM APIs: OpenAI (Responses API, function calling, Structured Outputs), Anthropic Claude (tool use), Azure OpenAI
Agents: LangGraph, LangChain, Model Context Protocol (MCP), multi-agent workflows, tool calling, human-in-the-loop
RAG and Retrieval: text-embedding-3, chunking, semantic search, OpenSearch, ChromaDB, prompt design, grounding
Evaluation and Safety: LangSmith, LLM-as-judge, faithfulness, tool-call accuracy, schema validation, PII redaction
Programming and APIs: Python, SQL, FastAPI, Flask, REST APIs, Kafka, MongoDB, Git
Cloud and MLOps: AWS (SageMaker, EKS, OpenSearch), Azure, Docker, GitHub Actions, CI/CD, MLflow, Snowflake, dbt
ML and NLP: PyTorch, Scikit-Learn, Hugging Face Transformers, BERT, spaCy, NLTK, intent classification, NER, A/B tests

WORK EXPERIENCE

AI Engineer | JPMorgan Chase & Co., New York, NY

Jan 2026 - Present

- Build a LangGraph multi-agent workflow on OpenAI and Anthropic Claude APIs that checks Chase for Business campaign copy against UDAAP-aligned disclosure rules, product terms, and brand guidelines, cutting review rework by about 30%.
- Expose the approved disclosure library, offer terms, and brand standards to the agents as MCP servers with read-only tool scopes, and ground each check in text-embedding-3 retrieval over previously cleared campaigns in Amazon OpenSearch.
- Score campaigns for launch risk with a PyTorch classifier trained on historical approval outcomes and route high-risk items to human reviewers first, holding 85% precision and saving the reviewers about 6 hours a week.
- Gate every prompt and agent change on a LangSmith evaluation suite of reviewer-labeled examples, scoring faithfulness and tool-call accuracy with an LLM judge plus JSON schema and PII redaction checks before Model Risk review.
- Deploying the service on Amazon SageMaker and EKS behind FastAPI, wired to campaign authoring and execution systems over REST APIs and Kafka, handling about 2,000 checks a day at under 300 ms latency with full tracing and audit logs.
- Partner with Marketing, Legal, Compliance, and Controls to turn recurring review findings into pre-launch rules.

Jr. AI Engineer | InfoWay Solutions, Mumbai, India

Mar 2021 - Aug 2024

- Began with the data work for a private bank client's complaint-routing model, cleaning and labeling net banking, UPI, and card-dispute tickets with spaCy and NLTK and training Scikit-Learn baselines by issue type.
- Fine-tuned Hugging Face BERT models on those tickets for intent classification and extraction of account, transaction, and loan references, lifting F1 from 0.78 to 0.89 and cutting manual routing time by about 40%.
- Moved reference extraction from regex to GPT-3.5 Turbo function calling, returning validated JSON to the CRM.
- Took the bank's internal help-desk assistant from prototype to rollout, chaining LangChain retrieval over ChromaDB embeddings of product FAQs, SOPs, and RBI circulars with GPT-4 on Azure OpenAI and keeping chat history in MongoDB.
- Served the models and the assistant as FastAPI and Flask services in Docker on Azure through GitHub Actions pipelines, cutting release time from two days to four hours, with each version tracked in MLflow and A/B tested first.

ACADEMIC PROJECTS

RAG Legal Document Assistant (Python, LangChain, Hugging Face, ChromaDB)

- Created a legal document Q&A tool on local 384-dimension Hugging Face embeddings, so it ran without any paid API calls.
- Indexed vectors in ChromaDB and tuned recursive chunking to return context in under a second from 400+ page documents.

Airbnb ML Feature Store (Snowflake, dbt Core, Python)

- Modeled Bronze, Silver, and Gold layers in Snowflake with dbt Core, using incremental loads to keep compute costs down.

EDUCATION

Stevens Institute of Technology

Master of Science in Computer Science

Hoboken, NJ

Sep 2024 - May 2026

Somaiya Vidyavihar University

Bachelor of Science in Information Technology (Honors)

Mumbai, India

Jul 2020 - Apr 2023

CERTIFICATIONS

[AWS Certified Data Engineer - Associate](#) | [AWS Academy Data Engineering](#)