

Vedant Somani

vedants0@uw.edu | +1 (740) 760 0011 | Seattle, WA | [LinkedIn: vedant-somani17](#) | [GitHub: vedant1711](#) | [Portfolio](#)

EDUCATION

University of Washington, Seattle
Master of Science - Information Management

June 2027
GPA: 4.0

Acropolis Institute of Technology and Research, India
B.Tech. - Computer Science Engineering

June 2025

WORK EXPERIENCE

Text Mining Specialist, UW Libraries & eScience Institute, (NLP, AI Research, RAG, Teaching)

Mar 2026 – Present

- Ingested **50K+** of a 400K-article target into **Open IRE**, an automated pipeline I built to collect, rights-classify, and preserve UW-authored articles, owning the database schema and system architecture.
- Recovered **6,800+** missing documents by building a custom exponential **backoff retry middleware** and fixing a self-duplicate dedup bug behind a scraping pipeline failure, validated with a 295-test e2e suite.
- Improved **recall@5** by 14pp on a 500-query eval set by redesigning **RAG** chunking from fixed 512-token to a sentence-anchored sliding window, 20% overlap; shipped a **Haystack**-based service cutting manual research query time **~30%**.

Fullstack Developer Intern, OS Technology Service, (React Native, Django, PostgreSQL)

Jan 2025 – Jul 2025

- Supported **500+ concurrent users** per event over **WebSocket** at **~200ms latency** and lifted session length **25%**, engineering a cross-platform app (My Dj Live) with real-time engagement in **React Native** and a **Django REST API**.
- Held **99.9%** uptime and **99.5%** crash-free sessions while cutting deploy time **40%**, deploying **AWS** infrastructure (**EC2**, **Fargate**, **CloudWatch**) with **CI/CD automation**, OTA updates, and analytics (**Crashlytics**, **DataDog**).

Machine Learning Intern, Xenon Nation, (PyTorch, FastAPI, Docker)

Mar 2024 – Sept 2024

- Cut inference **latency 35%** (~250ms) by exporting models to **ONNX** and adding response caching in **FastAPI** microservices deployed via **Docker** on AWS ECS.
- Boosted sentiment **F1-score 18%** (74% to 87%) fine-tuning **cardiffnlp/twitter-roberta-base-sentiment** via custom **PyTorch** loop; reaching **89%** precision, **85%** recall on review data.

PROJECTS

Vegapunk – Autonomous AI Code Agent, (LangGraph, FastAPI, NVIDIA NIM, Next.js, Docker)

- Improved bug-file localization 22% to 83% in an autonomous agent that turns a GitHub issue into a pull request, fusing **BM25** with a **tree-sitter code graph**, exposed via **MCP** server to **Claude Code**, **Cursor**, and **Cline**.
- Cut prompt tokens 31%** in the 7-node **LangGraph** pipeline behind it: K parallel patches from local **Ollama** or cloud **LLMs**, tested in isolated git worktrees, streamed over SSE from **FastAPI** to **Next.js**.

AgentGate, Open-source statistical CI gate for LLM agents, (LLM Evaluation, OpenTelemetry, SciPy)

- Cut false regression alarms from **~87% to 5%** in a **CI gate** that blocks a pull request when an **AI agent** has statistically regressed using **paired significance tests** and FDR correction across 42 metrics.
- Exposed a 9× gap between how coherent a model sounds (0.99) and how often it succeeds (0.11) by benchmarking two **LLM agents** across 666 runs and 4.5M tokens on **τ²-bench**, using **LiteLLM**, **DuckDB** and **OpenTelemetry** tracing.

ResolveIQ – AI-Powered Incident Knowledge Management Platform, (LangChain, FastAPI, Pinecone)

- Eliminated 90% of manual documentation overhead by architecting an event-driven microservices platform that auto-generates structured **Confluence KB articles** from resolved **Jira tickets**, aggregating **Slack** and **Zoom** context.
- Cut **MTTR** for recurring issues by an estimated **30%** with a semantic retrieval engine (**Voyage AI embeddings**, **Pinecone vector search**, **GPT-4o**) that surfaces relevant KB articles inline in Jira.

SKILLS

Languages and Data: C#, C++, Python, Java, JavaScript, TypeScript, MySQL, PostgreSQL, MongoDB, Pinecone

AI & Automation: Claude Code, Cursor, Github Copilot, GPT-4o, Langchain, LangGraph, LangSmith, LlamaIndex, RAG pipelines, BERT, Haystack, AutoGen, CrewAI, MCP, Tool calling, Multi Agent systems, LLM Evals, RAGAS, Phoenix, DeepEval

Machine Learning & Data Systems: PyTorch, TensorFlow, Scikit-learn, NumPy, Pandas, Snowflake, Hadoop, Scala, Apache Spark, PySpark, Apache Kafka, OpenCV, MATLAB, XGBoost, Supervised Learning

Platforms: AWS, Docker, Linux/Unix, Kubernetes, Git & GitHub, FastAPI, Django, React, Next.js, Node.js, Postman, Azure

ACHIEVEMENTS & PUBLICATIONS

- Instructed at the **AI in Practice Summer Institute**
- Dempsey Startup Competition 2026 **Investment Round Finalist**
- iEngage – Salesforce – Lovable **Hackathon Runner-up**
- Association of Information Management Students at UW (UW AIMS) **Product Competition Runner-up**
- IEEE ACROSET 2025: PyFao-Driven Precision Irrigation and Fertilizer Optimization in Agriculture**
- IEEE ICTBIG 2024: Elderly Well-Being Through Volunteer Engagement**
- AWS Academy Graduate – AWS Academy Cloud Architecting**