

ABHAY MADALA

Tampa, FL | (656) 223-9669 | madalaabhay1@gmail.com | abhaymadala@usf.edu | Permanent Resident
github.com/ABHAY-MADALA | linkedin.com/in/abhay-madala-8855742a9

PROFILE

Computer Science student building agentic AI systems, RAG pipelines, LLM evaluation workflows, backend APIs, and tool-integrated applications. Hands-on experience with document parsing, vector databases, prompt engineering, retrieval evaluation, and API-driven AI application development.

EDUCATION

University of South Florida - Tampa, FL

Expected Graduation: 2028

Bachelor of Science in Computer Science, Minor in Mathematics

Relevant Coursework: Data Structures, Program Design, Computer Organization, Discrete Structures, Calculus I-III, Physics I-II

In Progress: Software Engineering, Operating Systems, Computer Architecture & Lab, Vector Calculus

EXPERIENCE

AI Intern | Evoke Technologies

July 2026 - August 2026

- Developed Retrieval-Augmented Generation applications using Large Language Models and OpenAI APIs.
- Designed document processing pipelines for parsing, chunking, embedding generation, vector storage, and semantic retrieval.
- Researched and implemented prompt engineering techniques to improve response quality, retrieval accuracy, and LLM performance.
- Evaluated AI systems with RAGAS metrics including Faithfulness, Answer Relevancy, Context Precision, and Context Recall.
- Collaborated with mentors to explore RAG architecture and evaluation workflows.

PROJECTS

AgentHub Studio - Agentic AI Workspace

Python | FastAPI | LangGraph | OpenAI | SQLite | ChromaDB

- Built a local agentic AI workspace that routes user requests across SQL databases, document retrieval, Gmail, Calendar, and other tools.
- Designed multi-step workflows for tool selection, database querying, retrieval, and natural-language response generation.
- Implemented approval gates for consequential actions and structured agent/tool orchestration.
- Integrated FastAPI, SQLite, ChromaDB, OpenAI APIs, and Google OAuth/API services.

Enterprise Document Intelligence & RAG Evaluation

Python | OpenAI | Docling | ChromaDB | Gradio | RAGAS

- Built an end-to-end RAG chatbot capable of processing multiple document formats using Docling.
- Implemented document parsing, chunking, embedding generation, semantic retrieval, ChromaDB vector storage, and LLM response generation.
- Integrated RAGAS to measure Faithfulness, Answer Relevancy, Context Precision, and Context Recall.
- Developed an interactive Gradio interface for document upload and conversational querying.

TECHNICAL SKILLS

Programming Languages: Python, C++

Web & Backend: HTML, FastAPI, REST APIs

AI & LLM: Retrieval-Augmented Generation (RAG), Agentic Workflows, RAGAS, Prompt Engineering, LLM Evaluation, Embeddings, Vector Databases

Frameworks & Libraries: LangGraph, Gradio, Docling, ChromaDB

APIs & Data: OpenAI API, Gemini API, Google APIs, SQLite, JSON

Developer Tools: Git, VS Code

Core Concepts: Document Processing, Semantic Search, Information Retrieval, Chunking