

Sugam Panthi

Sugam.Panthi@usm.edu — 601-307-0189 — sugampanthi.com.np
github.com/vein05 — linkedin.com/in/sugam-panthi

Core Competencies

AI/ML: LLM Applications, Conversational Memory Systems, RAG Pipelines, VLMs, Predictive Modeling, Fine-tuning
Architecture & Infrastructure: AWS (ECS, Fargate, Lambda, ECR), Docker, Terraform, CI/CD, Microservices
Frameworks & Tools: Go, Python, LangChain, PyTorch, Pinecone, Flask, React, C/C++

Industry Experience

Co-Founder & CTO, *MagnoliaEd LLC*

Apr 2025 – Present

- Built and shipped an AI-powered education platform with personalized instructor chatbots and role-based dashboards; deployed across multiple college classrooms.
- Architected a hybrid monolithic system on AWS Fargate using Docker and Terraform, scaling chat, analytics, and auth independently across 1,000+ concurrent sessions.
- Established CI/CD pipelines with GitHub Actions, cutting deployment friction and automating test coverage across the full stack.
- Secured pre-seed funding through Co-Builders powered by Microsoft; led investor outreach and placed in multiple pitch competitions.

AI & ML Intern, *Prediction 3D*

May 2025 – August 2025

- Developed predictive models in PyTorch on multimodal construction-site data, combining images and sensor readings to generate safety and productivity insights for field teams.
- Integrated AI features into client-facing applications with LangChain and Pinecone, collaborating within a 20-engineer team to deploy ML capabilities into production workflows.
- Contributed to a microservice-based data pipeline separating ingestion, preprocessing, and model inference, improving maintainability and deployment flexibility.

AI Research Assistant, *University of Southern Mississippi*

Nov 2024 – Present

- Conducted research on conversational memory evaluation, analyzing how target definitions in transformed memory benchmarks can change retrieval conclusions even when ranked outputs remain fixed.
- Built reproducible evaluation analyses across long-horizon conversational benchmarks, studying representation-sensitive scoring behavior in systems that store raw turns, derived observations, and canonical facts.
- Researched and modeled U.S. plastic waste trends using LSTM networks, cleaning and visualizing longitudinal datasets and co-authoring a published CIB conference paper on plastic recycling.

Engagements & Recognition

- **Publication:** Same Ranking, Different Winner: How Scoring Targets Shape LLM Memory Benchmarks. Findings of EMNLP - arXiv:2605.24060
- **Preprint:** Fixed RAG Compression Collapses Measured Reader Scaling. Under review - arXiv:2606.21807
- **Preprint:** Outcome Monitors: Recovery Affordances for Silent Tool Failures. Under review - arXiv:2608.19303
- **Pitch Competition Winner**, Mississippi Polymer Institute May 2025
- **Hackathon Winner**, Davidson College January 2025
- **Hackathon Winner**, University of Southern Mississippi November 2024

Projects

Pali (Link) — Open-Source Memory Infrastructure for LLMs

- Local-first, inspectable memory runtime for LLM agents; exposes REST, MCP, and dashboard interfaces over a shared memory core written in Go.
- Multi-tenant storage and retrieval with lexical + vector search, candidate fusion, and configurable embedding/scoring providers.
- Pluggable backends (SQLite, Qdrant, Neo4j); benchmarked across LoCoMo, LongMem, and BEAM.

deepseek-go (Link) — Open-Source Go SDK

- Most popular Go client for DeepSeek-compatible APIs; 325 stars, 38 forks, 45+ issues closed, 30+ PRs.
- Shipped structured releases through v1.3.3 covering streaming support, reliability fixes, and contributor onboarding.
- Maintained issue triage and community contributions; used in production by multiple companies.

Cohesion — AI-Powered SQL Assistant (Link) — Hackathon Winner

- Built an AI-driven schema generation tool with a custom speed-optimized CSV format; implemented backend services using Go, Gin, and Docker.

Education

The University of Southern Mississippi, Honors College
B.S. in Computer Science

Hattiesburg, MS
GPA: 4.0 / 4.0