

SUGAM PANTHI

Undergraduate Researcher in Conversational Memory & LLM Evaluation
Sugam.Panthi@usm.edu · 601-307-0189 · spanthi.com · github.com/vein05 ·
linkedin.com/in/sugam-panthi · Google Scholar

RESEARCH INTERESTS

LLM evaluation: how measurement choices decide what a benchmark concludes. Scoring targets in conversational memory, and compression that flattens measured reader scaling.

Tool-using agents: reliability under silent tool failure, and recovery affordances for repairing corrupted results.

LLMs and human input: unmarked seams in a prompt, including the boundary between pasted material and user speech.

EDUCATION

The University of Southern Mississippi, Honors College Aug 2024 – May 2028 (Expected)
B.S. in Computer Science; GPA: 4.0/4.0 *Hattiesburg, MS, USA*

PUBLICATIONS

Peer-reviewed

- [C1] **S. Panthi** and R. Abdelfattah, “Same Ranking, Different Winner: How Scoring Targets Shape LLM Memory Benchmarks,” *Findings of EMNLP*, 2026. arXiv:2605.24060
- [C2] **S. Panthi** and F. Zhang, “A Comprehensive Review of Plastic Recycling in the Construction Industry: Challenges and Opportunities in the U.S.,” in *CIB Conferences*, vol. 1, art. 63, 2025. Purdue e-Pubs

Under review

- [U1] **S. Panthi** and R. Abdelfattah, “Fixed RAG Compression Collapses Measured Reader Scaling,” 2026. arXiv:2606.21807
- [U2] **S. Panthi** and R. Abdelfattah, “Outcome Monitors: Recovery Affordances for Silent Tool Failures,” 2026. arXiv:2608.19303

RESEARCH EXPERIENCE

Scoring-Target Non-Invariance in LLM Memory Benchmarks [C1, U1] Nov 2025 – Present
Lead Undergraduate Researcher; Advisor: Rabab Abdelfattah

- Found that LLM memory benchmarks score the same retrieved item as a hit or a miss depending only on which version of a fact they accept as correct. Named the problem target non-invariance: scores, system rankings, and design recommendations all move while the retrieved output stays fixed.
- Built an audit that rescores saved retrieval outputs under three definitions of a correct answer (Raw, Source, Canonical) without rerunning retrieval. On LoCoMo and LongMemEval-S the definition alone changes nDCG on 83–94% of shared queries and reverses system rankings on Mem0 and MemoryOS.

Outcome Monitors: Recovery Affordances for Silent Tool Failures [U2] 2026
Lead Undergraduate Researcher

- Built Outcome Monitors, which catch tool calls that fail silently in LLM agents. Each result is checked against rules learned from unrelated successful runs; when a rule breaks, the agent is told which one

and which tools can recover.

- In frozen, prespecified evaluations with injected failures, raised ToolMaze task completion from 10.9% to 28.1% across four models from two providers, and τ -bench retail completion by 14.0 and 12.0 points. Ablations show the list of recovery tools is what does the work.

SIEVE: Post-Retrieval Evidence Compilation

2026

Sole Author

- Tested four ways of adapting how retrieved evidence is compressed before the model reads it, under one controlled pipeline with success and failure criteria fixed in advance. Only conditioning on the kind of query survived.
- Built a query-conditioned compiler that sits between BM25 retrieval and the reader model and keeps its gains across readers, avoiding the 30–40 point drops that one fixed compression policy causes.

SEAM: Absorption at Unmarked Prompt Seams

Jul 2026 – Present

Lead Undergraduate Researcher

- “Can LLMs Separate Pasted Artifacts from User Speech? Absorption at Unmarked Prompt Seams”: measuring whether models distinguish pasted material from user-authored text when no delimiter marks the boundary, and where the seam is absorbed. Code: github.com/aimsresearchlab/seam.

Plastic-Waste Prediction with LSTMs [C2]

Nov 2024 – 2025

Undergraduate Research Assistant, The University of Southern Mississippi

- Modeled U.S. plastic-waste trends with LSTM networks; cleaned and visualized longitudinal datasets; authored methodology and results sections for two journal submissions.
- Built a React Native logistics app for recycling-plant waste transportation (Clerk, MongoDB, Mapbox).

INDUSTRY EXPERIENCE

Co-Founder & CTO, MagnoliaEd LLC

Apr 2025 – Present

- Built an AI-powered education platform with personalized instructor chatbots (RAG over course materials) and role-based dashboards; deployed across multiple college classrooms.
- Architected the system on AWS Fargate with Docker and Terraform, scaling chat, analytics, and auth across 1,000+ concurrent sessions; CI/CD with GitHub Actions.
- Press: Innovate Mississippi, Mississippi Today.

AI & ML Fellow, Prediction 3D

May 2025 – Aug 2025

- Developed predictive models in PyTorch on multimodal construction-site data (images + sensor readings); integrated AI features into client applications with LangChain and Pinecone within a 20-engineer team.

Backend Intern, Best One Network

Feb 2024 – May 2024

- Rebuilt the payment-confirmation API (Django → Flask/WSGI); remediated 2 critical bugs pre-release, reducing post-deployment issues by 10%.

HONORS & AWARDS

Pre-Seed Funding, Co-Builders powered by Microsoft

2025

- Secured \$27,000 pre-seed funding for MagnoliaEd; led investor outreach.

3rd Place & People's Choice, MPI Pitch Competition	May 2025
Hackathon Winner, Davidson College	Jan 2025
Hackathon Winner (1st Place), USM Hatchathon	Nov 2024

SELECTED PROJECTS & OPEN SOURCE

Pali: Open-Source Memory Infrastructure for LLMs GitHub

- Local-first, inspectable memory runtime for LLM agents in Go; REST, MCP, and dashboard interfaces over a shared memory core; multi-tenant lexical + vector retrieval with candidate fusion.
- Pluggable backends (SQLite, Qdrant, Neo4j); benchmarked on LoCoMo, LongMemEval, and BEAM.

deepseek-go: Go SDK for DeepSeek-Compatible APIs GitHub

- Lead maintainer of the most popular Go DeepSeek client (345 stars, 35 forks), covering the full DeepSeek API surface; 45+ issues closed, 30+ PRs merged, releases through v1.3.3; used in production by multiple companies.

Cohesion: AI-Powered SQL Assistant (Hackathon Winner) GitHub

- AI-driven schema generation with a custom speed-optimized CSV format (200% faster mock-data analysis); Go/Gin/Docker backend.

nomnom: GenAI Bulk File Renaming & Organization GitHub

- CLI tool in Go for bulk renaming and organizing files with LLMs; distributed via a Homebrew tap.

LEADERSHIP & SERVICE

Head of Entrepreneurship Initiatives 2025 – Present

Google Developer Group on Campus, The University of Southern Mississippi *Hattiesburg, MS*

- Led the chapter's Semester Meetups, one each semester, programming sessions on entrepreneurship and technical founding for undergraduate attendees.
- Co-organized three further GDG events, including a Robo Soccer tournament and chapter seminars.
- Mentored two undergraduates on early-stage software projects and research directions.

Executive Officer and Spokesperson 2025 – 2026

Nepalese Students Association, The University of Southern Mississippi *Hattiesburg, MS*

- Organized cultural and community events drawing 300+ participants, covering programming, logistics, and volunteer coordination.
- Served as spokesperson for the association, representing it to university administration and the wider campus community.

TECHNICAL SKILLS

Languages: Python, Go, C/C++, TypeScript/JavaScript, SQL

ML/NLP: PyTorch, Hugging Face Transformers, LangChain, RAG pipelines, LLM/IR evaluation (nDCG, benchmark design), fine-tuning, vector search (Pinecone, Qdrant)

Infrastructure: AWS (ECS, Fargate, Lambda), Docker, Terraform, PostgreSQL, MongoDB, Neo4j, CI/CD, Linux

Tools: Git, L^AT_EX, tmux