

Neeraj Venna

908-391-4484 | nvenna@ucsd.edu | [linkedin.com/in/neeraj-venna](https://www.linkedin.com/in/neeraj-venna) | github.com/kvitkas

Research Interests: World models, agentic memory, partially observable agents, model-based reasoning, uncertainty, continual adaptation

EDUCATION

University of California, San Diego
B.S. Mathematics-Computer Science

Expected 12/2027
La Jolla, CA

RESEARCH EXPERIENCE

AI Fellow

Basis Set Ventures

Summer 2026

San Francisco, CA

- Worked with Basis Set portfolio companies including Mem0 to benchmark agent memory on long-running, multi-session tasks involving temporal recall, conflicting updates, irrelevant memories, and paraphrased queries
- Designed controlled evaluations to distinguish failures in memory retrieval from failures in downstream reasoning, rather than treating end-task accuracy as a single aggregate metric
- Analyzed retrieval outputs and execution traces across memory configurations, comparing accuracy, context usage, latency, and characteristic failure modes

AI Research Intern

Salk Institute of Biological Studies

09/2024 – 09/2025

La Jolla, CA

- Developed Python evaluation tooling with PyTorch, OpenCV, NumPy, and pandas for long-timescale behavioral models, automating pose quality control, feature extraction, and failure analysis across large experimental datasets
- Built Dockerized multiprocessing and SLURM inference pipelines for large-scale pose data and downstream behavioral analysis
- Improved tracking reliability by 28% through validation workflows targeting occlusions, rare behaviors, and other high-error model cases
- Investigated model failures over long behavioral sequences where tracking or classification errors could propagate into downstream measurements

SELECTED RESEARCH

BeliefSpec

Predictive Memory for Partially Observable Agents

2026

| *Python, PyTorch, MiniGrid*

- Studied whether learning to predict future observations produces internal memory representations that are more useful for decision-making under partial observability
- Designed a controlled delayed-clue environment where an agent observes information early in an episode, loses direct access to it, and must later use memory to make the correct decision
- Compared six matched memory conditions: current observation, recent history, structured memory, episodic retrieval, task-only recurrent memory, and recurrent memory trained jointly with action-conditioned next-observation prediction
- Task-only GRU reached 100% observed-clue decision accuracy across all three pilot seeds, while predictive training reached 50%, 100%, and 100%; all methods remained at chance when the clue was never observed, providing a control against hidden-information leakage
- Audited 6,336 paired post-clue prediction targets and found they were invariant to clue identity, revealing that the auxiliary prediction task did not necessarily require preservation of the task-relevant latent state
- Used this failure to motivate a sharper causal decomposition of agent errors: information never observed, information observed but forgotten, and information retained but not used correctly by the downstream policy

Research Memory

Long-Term Memory for Scientific Work

2026

| *Python, LlamaIndex, ChromaDB, Llama 3.2, n8n, Docker*

- Built a local research-memory system after repeatedly remembering conclusions from papers while losing the original experiment, passage, or note where I encountered them
- Indexed papers, notes, and experiment logs with source, section, date, and topic metadata so retrieved answers preserve the underlying evidence rather than relying only on model-generated summaries
- Designed retrieval around provenance: answers surface the original supporting passages so claims can be checked against source material before being incorporated into new reasoning
- Automated document ingestion and indexing with n8n and Docker, allowing the memory store to grow continuously as new papers and experimental notes are added
- Kept retrieval and generation local for private research material while using the project to explore how persistent memory systems handle accumulated knowledge over time

ADDITIONAL RESEARCH EXPERIENCE

Research Assistant

11/2024 – 11/2025

Scripps Research

La Jolla, CA

- Worked on brainstem interoception experiments studying neural representations of internal physiological signals
- Performed ROI extraction and quantitative analysis for two-photon imaging data using Cellpose and FIJI
- Helped develop analysis workflows for converting imaging outputs into interpretable measurements of neural population responses

Data Engineer Intern

09/2023 – 05/2024

Network Contagion Research Institute

New Brunswick, NJ

- Built a React, FastAPI, and PostgreSQL platform for collecting, querying, and analyzing structured incident reports
- Designed REST APIs and relational database schemas supporting report submission, filtering, retrieval, and reproducible downstream analysis
- Developed NLP and geospatial pipelines with scikit-learn, spaCy, pandas, and GeoPandas to identify recurring themes, geographic hotspots, and temporal trends

RESEARCH DIRECTION

- Interested in agents that learn compact predictive models of their environment and use those models to reason, plan, and adapt under uncertainty
- Particularly interested in when persistent memory helps or hurts an agent, how models determine which information should survive over long horizons, and how memory interacts with downstream reasoning
- Interested in experimental settings that separate perception, memory, prediction, and planning failures rather than evaluating only aggregate task success

TECHNICAL SKILLS

Programming: Python, C++, TypeScript/JavaScript, SQL

ML / Research: PyTorch, MiniGrid, NumPy, pandas, OpenCV, scikit-learn, recurrent models, predictive modeling, experiment design, ablation studies, model evaluation

Agents / Memory: LangGraph, LangChain, LlamaIndex, ChromaDB, RAG, embeddings, retrieval, tool calling, agent memory, agent evaluation

Systems: Docker, Git, SLURM, multiprocessing