Sai Vikram Kolasani

203-997-7329 | saikolasani@berkeley.edu | saikolasani.github.io | www.linkedin.com/in/saikolasani

EDUCATION

University of California, Berkeley **Berkeley, CA**

Bachelor of Arts in Statistics, Computer Science, Public Policy Minor *May 2025*

- **Cumulative GPA**: 3.67/4.0; **ACT**: 35.25/36, **SAT:** 1520/1600
- **Relevant Coursework:** Data Structures & Algorithms, Techniques of Data Science, Linear Algebra & Differential Equations, Multivariable Calculus, Probability, Game Theory, Machine Learning, Artificial Intelligence, Computer Vision
- **Activities**: Alpha Kappa Psi, Poker @ Berkeley, Capital Investments @ Berkeley, Generative AI @ Berkeley

Awards & Honors: We The People Competitive Debator (1st-state, 1st-northeast, 9th-national, Received Citation of Excellence from CT government), Venture x Nexus Case Competition (Finalist), Ribbon of Excellence for Research at Dan Lab, MathWorks Math Modeling Challenge (Honorable Mention), Dean's List (Fall 2022), Datastax Internal Langflow Hackathon (Winner)

RESEARCH EXPERIENCE

Sky Computing Lab @ UC Berkeley **Berkeley, CA**

Artificial Intelligence and Machine Learning Researcher *Jan 2024 - Present*

- Co-released **Agent Arena**, an interactive sandbox where users can compare, visualize, and rate agentic workflows personalized to their needs, along with Shishir Patil, Joseph Gonzalez, and Ion Stoica
- Spearheading the development of **Ember**, an open-source compositional framework for constructing **compound AI systems and "Networks of Networks" (NoNs)**. Ember combines the structural and efficiency benefits of JAX/XLA with the compositional user experience of PyTorch/FLAX, enabling the creation of modular, scalable AI workflows
- Developing methods to prevent hallucinations in AI systems by ensemble approaches across models with **Jared Davis and Matei Zaharia**

Long-Term AI Safety Research @ Cornell University **Berkeley, CA**

Artificial Intelligence and Machine Learning Researcher *Jun 2024 - Present*

- Developed and fine-tuned AI models using reinforcement learning with human feedback (RLHF) to incorporate real-time emotion-based signals from facial emotion recognition (FER) and Valence-Arousal-Dominance (VAD) metrics, enhancing empathetic and personalized interactions
- Designed and implemented an evaluation framework combining human experience ratings and emotional state analyses, demonstrating significant improvements in AI responses to negative emotional states, with applications in mental health support and conversational safety

UC Berkeley Sociology Department **Berkeley, CA**

Natural Language Processing Research Assistant *Jun 2023 - May 2024*

- Worked with Professor Heather A. Haveman to analyze the topics that US workers bring up when describing their jobs and workplaces. Conducting an inductive study using topic models based on word embeddings from Google's BERT
- Built a hybrid Fixed-Effect model to estimate the effects of time-invariant(ownership and size category) and time-varying predictors(age, managerial job indicator, and technical job indicator) on firm ratings

Dan Lab @ UC Berkeley Department of Molecular & Cell Biology **Berkeley, CA**

Machine Learning Research Assistant *Jan 2023 - May 2023*

- Constructed deep learning models to achieve unsupervised behavioral classification of laboratory mice across all sleep and active stages, used deep learning models to analyze the activation of neurons across various sleep and active stages. Achieved 94% accuracy with behavior classification

SELECTED PUBLICATIONS

LLM CHESS: Benchmarking Reasoning and Instruction-Following in LLMs through Chess

- **Publication:** Kolasani, Sai, et al. (2025). LLM CHESS: Benchmarking reasoning and instruction-following in LLMs through chess. **NeurIPS 2025**: First Workshop on Foundations of Reasoning in Language Models. **(First Author)**

Gorilla X LMSYS Agent Arena

- **Publication:** Yekollu, Nithik, et al. "Agent Arena: A Platform for Evaluating and Comparing LLM Agents." Gorilla, University of California, Berkeley, https://gorilla.cs.berkeley.edu/blogs/14_agent_arena.html. (**Co-first Author**)

Ember: An inference-time scaling architecture framework.

- **Publication:** Davis, J. Q., et al. "Ember: An inference-time scaling architecture framework." https://github.com/PyEmber/ember **(Co-first Author)**

Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks

- **Publication:** Kolasani, S.V. and Assaf, R. (2020) Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. Journal of Data Analysis and Information Processing, 8, 309-319. https://doi.org/10.4236/jdaip.2020.84018 **(First Author)**

PROFESSIONAL EXPERIENCE

Doordash **San Francisco, CA**

Machine Learning Engineer Intern *May 2025 - Present*

- Designed and prototyped an **end-to-end GenAI recommendation system** to enhance user engagement and conversion on store pages, targeting a 20%+ uplift by surfacing contextually relevant items.
- Built and deployed a context-aware ranking pipeline utilizing large language models (LLMs) and embeddings to interpret user signals, retrieve top-k recommendations, and generate evidence-based match scores.
- Engineered a dynamic carousel assembly feature to intuitively present ranked recommendations with automatically generated titles and supporting evidence in a user-friendly horizontal layout.
- Created an adaptive real-time user preference model, enabling personalized recommendations by continuously updating user-defined attributes such as cuisine types and spice tolerance.

Value Buddy **San Francisco, CA**

Artificial Intelligence Engineer *Feb 2025 - April 2025*

- Engineered a dynamic multi-agent orchestration system using **AG2** and **RAG** frameworks to automatically generate custom-trained chatbots for new reports, ensuring isolated, report-specific data access and real-time insights.
- Designed and integrated a master orchestrator agent to intelligently route queries to specialized sub-agents across multiple persistent vector databases and APIs, delivering targeted, comprehensive financial analytics for banking clients.

Arize AI **Berkeley, CA**

Software Engineer Intern *Jun 2024 - Aug 2024*

- Enhanced Efficiency and Maintainability of the **demo models system** by refactoring and modularizing the demo models, reducing code complexity and eliminating duplication, resulting in a more maintainable and scalable architecture
- Implemented automated updates and improved reliability by adding cronjobs for automated model updates and standardized data processing, leading to a **30% reduction in manual intervention** and improved system reliability
- Implemented **auto instrumentation for GuardrailsAI** to allow collecting traces generated via OpenInference instrumentation. This will enable traces for any guards generated by GuardrailsAI to be analyzed in Arize Phoenix

DataStax (IBM) **Santa Clara, CA**

Software Engineer Intern *Jan 2024 - Jun 2024*

- Spearheaded the creation of **RAGulate,** a new module for RAGStack, enhancing the evaluation and performance monitoring of RAG pipelines by introducing comprehensive metrics for answer correctness, relevance, and groundedness
- Played a crucial role in building and refining features in the first enterprise release of **Langflow** by adding Astra DB's vectorize capability, allowing embedding generation at the database layer, leaving developers to focus on application logic

Priceline **Norwalk, CT**

Machine Learning Engineer Intern *Jun 2023 - Aug 2023*

- Designed and implemented a **travel destination recommendation system** by leveraging clustering algorithms such as **DBSCAN** and **BERTopic** to group similar travel destinations. Integrated the system with user preference data and past purchase history to dynamically recommend personalized travel options and amenities, improving user engagement.
- Conducted extensive prompt engineering to optimize the performance of **Penny**, Priceline's Gen AI Assistant, utilizing Vertex AI and Generative AI frameworks. Implemented tailored conversation flows and contextual prompts, resulting in a measurable improvement in user satisfaction and task completion rates.

SKILLS AND INTERESTS

Languages: Python, C++, C, Golang, R, SQL, Scheme, Java, HTML, CSS, JavaScript
Technologies: NumPy, Pandas, Scikit-learn, NLTK, Keras, Tensorflow, Pytorch, Google Cloud (GCP), Bazel, Kubernetes
Interests: Golf, Poker, Basketball, Tennis, Video Games, R&B music, Longboard, Gym, Anime, Pickleball, Boston Celtics