Shreya Shankar

shreyashankar@berkeley.edu | sh-reya.com | Google Scholar

Education

2021–26	**Ph.D.**, EECS, University of California, Berkeley Advisor: Aditya G. Parameswaran
2018–20	**M.S.**, Computer Science (part-time), Stanford University
2015–19	**B.S.**, Computer Science, Stanford University Advisor: Pat Hanrahan

Research Experience

Jul. 2027–	**Assistant Professor** (tenure-track), Carnegie Mellon University, Computer Science Department; by courtesy, Human-Computer Interaction Institute (HCII)
Jul. 2026–27	**Visiting Assistant Professor**, Carnegie Mellon University, Computer Science Department; by courtesy, Human-Computer Interaction Institute (HCII)
2021–2026	**PhD Researcher**, UC Berkeley, EECS — Data Systems & Foundations Group Areas: data management, AI-powered data systems, human-computer interaction.
2025	**Visiting Student Researcher**, Google Systems Research Group (Sunnyvale, CA, USA) Built a new optimizer for DocETL.
2022	**Research Intern**, Meta — AI Data Infrastructure Team (Menlo Park, CA, USA) Automatic data validation for ML pipelines.
2017–19	**Research Intern**, Google Brain (Mountain View, CA, USA) Mentors: Alex Kurakin and Ian Goodfellow; ML security and adversarial examples.

Awards & Honors

2026	Received "Best of ACM SIGMOD" recognition for our paper
2026	Received ACM CHI Best Paper
2026	Laude Slingshot Grant Recipient
2025	Selected for EECS Rising Stars Workshop
2025	Received ACM UIST Best Paper Honorable Mention
2025	Bridgewater Research Fellowship Recipient
2025	Mercor PhD Fellowship Finalist
2025	Berkeley Nominee (1 in 5), Apple AI/ML PhD Fellowship
2024	Berkeley EECS Evergreen Mentorship Award Recipient
2023	Heidelberg Laureate Forum Young Scholar
2022	NDSEG Fellowship Recipient
2022	Hertz Foundation Fellowship Finalist
2022	P.D. Soros Fellowship Finalist
2021	UC Berkeley EECS Excellence Award Recipient

Publications

Under Submission

P1. Ma*, R., **Shankar, Shreya***, Chen, R., Lin, Y., Zeighami, S., Ghosh, R., Gupta, A., Gupta, A., Gopal, T. & Parameswaran, A. G. *Can AI Agents Answer Your Data Questions? A Benchmark for Data Agents* 2026. https://arxiv.org/abs/2603.20576.

Conditionally Accepted or Under Revision

R1. Chopra, B., Chen, M., Dang, R., Park, C., **Shankar, Shreya**, Zeighami, S., Hartmann, B. & Parameswaran, A. G. *Who's Keeping Score? Interactive Steering of LLM-Powered Scoring with ATTUNE* Conditionally accepted to UIST 2026.

R2. Sun*, Y., Zeighami*, S., Chopra, B., **Shankar, Shreya** & Parameswaran, A. G. *Semantic Data Processing with Holistic Data Understanding* Under revision at VLDB 2027. https://arxiv.org/abs/2604.02655.

Peer-reviewed Conference Proceedings

C1. Lauro*, Q. R., **Shankar, Shreya***, Zeighami, S. & Parameswaran, A. *RAG Without the Lag: Interactive Debugging for Retrieval-Augmented Generation Pipelines* in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)* ⋆ **Best Paper** (2026). https://arxiv.org/abs/2504.13587.

C2. Liu, S., Ponnapalli, S., **Shankar, Shreya**, Zeighami, S., Zhu, A., Agarwal, S., Chen, R., Suwito, S., Yuan, S., Stoica, I., Zaharia, M., Cheung, A., Crooks, N., Gonzalez, J. E. & Parameswaran, A. G. *Supporting Our AI Overlords: Redesigning Data Systems to be Agent-First* in *Conference on Innovative Data Systems Research (CIDR)* (Chaminade, CA, USA, Jan. 2026).

C3. **Shankar, Shreya**, Zeighami, S. & Parameswaran, A. *Task Cascades for Efficient Unstructured Data Processing* in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* (ACM, New York, NY, USA, May 2026).

C4. Wei*, L. L., **Shankar, Shreya***, Zeighami, S., Chung, Y., Ozcan, F. & Parameswaran, A. G. *Multi-Objective Agentic Rewrites for Unstructured Data Processing* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2026* (2026). https://arxiv.org/abs/2512.02289.

C5. Zeighami, S., **Shankar, Shreya** & Parameswaran, A. *Cut Costs, Not Accuracy: LLM-Powered Data Processing with Guarantees* in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* ⋆ **Invited to "Best of SIGMOD"** (ACM, New York, NY, USA, May 2026).

C6. Zeighami, S., **Shankar, Shreya** & Parameswaran, A. *Featurized-Decomposition Join: Low-Cost Semantic Joins with Guarantees* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2026* (2026). https://arxiv.org/abs/2512.05399.

C7. Lin, R., Chopra, B., Lin, W., **Shankar, Shreya**, Hulsebos, M. & Parameswaran, A. G. *Rethinking Dataset Discovery with DataScout* in *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (Association for Computing Machinery, 2025). ISBN: 9798400720376. https://doi.org/10.1145/3746059.3747727.

C8. Lin, Y., Hulsebos, M., Ma, R., **Shankar, Shreya**, Zeighami, S., Parameswaran, A. G. & Wu, E. *Querying Templatized Document Collections with Large Language Models* in *IEEE 41st International Conference on Data Engineering (ICDE)* (2025), 2422–2435.

C9. **Shankar, Shreya**, Chambers, T., Shah, T., Parameswaran, A. G. & Wu, E. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2025* **18** (2025), 3035–3048.

C10. **Shankar, Shreya***, Chopra*, B., Hasan, M., Lee, S., Hartmann, B., Hellerstein, J., Parameswaran, A. & Wu, E. *Steering Semantic Data Processing With DocWrangler* in *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* ⋆ **Best Paper Honorable Mention** (Association for Computing Machinery, New York, NY, USA, 2025). ISBN: 9798400720376. https://doi.org/10.1145/3746059.3747625.

C11. Vir*, R., **Shankar, Shreya***, Chase, H., Hinthorn, W. & Parameswaran, A. *PROMPTEVALS: A Dataset of Assertions and Guardrails for Custom Production Large Language Model Pipelines* in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (Association for Computational Linguistics, Albuquerque, New Mexico, Apr. 2025), 4225–4245. https://aclanthology.org/2025.naacl-long.213/.

C12. Parameswaran, A. G., **Shankar, Shreya**, Asawa, P., Jain, N. & Wang, Y. *Revisiting Prompt Engineering via Declarative Crowdsourcing* in *Conference on Innovative Data Systems Research (CIDR)* (Santa Cruz, CA, USA, 2024). https://www.cidrdb.org/cidr2024/papers/p67-parameswaran.pdf.

C13. **Shankar, Shreya**, Garcia, R., Hellerstein, J. M. & Parameswaran, A. G. *"We Have No Idea How Models will Behave in Production until Production": How Engineers Operationalize Machine Learning* in *Proceedings of the ACM on Human-Computer Interaction (CSCW)* **8** (Apr. 2024). https://doi.org/10.1145/3653697.

C14. **Shankar, Shreya**, Li, H., Asawa, P., Hulsebos, M., Lin, Y., Zamfirescu-Pereira, J., Chase, H., Fu-Hinthorn, W., Parameswaran, A. G. & Wu, E. *SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2024* **17** (2024), 4173–4186.

C15. **Shankar, Shreya** & Parameswaran, A. G. *Building Reactive Large Language Model Pipelines with Motion* in *Companion of the 2024 International Conference on Management of Data (SIGMOD/PODS '24)* (ACM, 2024), 520–523. https://doi.org/10.1145/3626246.3654734.

C16. **Shankar, Shreya**, Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A. & Arawjo, I. *Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences* in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST)* ⋆ **Most Cited Paper, UIST 2024** (2024), 1–14.

C17. **Shankar, Shreya**, Fawaz, L., Gyllstrom, K. & Parameswaran, A. *Automatic and Precise Data Validation for Machine Learning* in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)* (ACM, 2023), 2198–2207. https://doi.org/10.1145/3583780.3614786.

C18. **Shankar, Shreya**, Macke, S., Chasins, A., Head, A. & Parameswaran, A. *Bolt-on, Compact, and Rapid Program Slicing for Notebooks* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2022* **15** (2022), 4038–4047.

C19. **Shankar, Shreya** & Parameswaran, A. *Towards Observability for Production Machine Learning Pipelines* in *Proceedings of the VLDB Endowment (PVLDB) / VLDB 2022* **15** (2022), 4015–4022.

C20. Dathathri, S., Dvijotham, K., Kurakin, A., Raghunathan, A., Uesato, J., Bunel, R. R., **Shankar, Shreya**, Steinhardt, J., Goodfellow, I., Liang, P. S. & Kohli, P. *Enabling Certification of Verification-Agnostic Networks via Memory-Efficient Semidefinite Programming* in *Advances in Neural Information Processing Systems (NeurIPS)* **33** (2020), 5318–5331.

C21. Elsayed, G. F., **Shankar, Shreya**, Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I. & Sohl-Dickstein, J. *Adversarial Examples That Fool Both Computer Vision and Time-Limited Humans* in *Advances in Neural Information Processing Systems (NeurIPS)* (Montréal, Canada, 2018), 3914–3924.

Workshop Papers

W1. Hulsebos, M., Lin, W., **Shankar, Shreya** & Parameswaran, A. *It Took Longer than I was Expecting: Why is Dataset Search Still so Hard?* in *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics (HILDA) at SIGMOD* (2024), 1–4.

W2. **Shankar, Shreya**, Halpern, Y., Breck, E., Atwood, J., Wilson, J. & Sculley, D. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World* in *NeurIPS 2017 Workshop: Machine Learning for the Developing World* (2017).

Invited Talks

T1. *Building Effective Unstructured Data Systems* Keynote, DashSys Workshop at VLDB. Aug. 2026.

T2. *Building Effective Unstructured Data Systems* Keynote, AIDM Workshop at SIGMOD. June 2026.

T3. *Building Effective Unstructured Data Systems* Carnegie Mellon University Computer Science Department. Mar. 2026.

T4. *Building Effective Unstructured Data Systems* MIT Electrical Engineering and Computer Science. Mar. 2026.

T5. *Building Effective Unstructured Data Systems* Columbia University Computer Science. Mar. 2026.

T6. *Building Effective Unstructured Data Systems* University of Washington Computer Science and Engineering. Feb. 2026.

T7. *Building Effective Unstructured Data Systems* NYU Computer Science. Feb. 2026.

T8. *Building Effective Unstructured Data Systems* University of Chicago Department of Computer Science. Feb. 2026.

T9. *Building Effective Unstructured Data Systems* UT Austin Computer Science. Jan. 2026.

T10. *AI-Powered Data Processing with DocETL and DocWrangler* NorCal Public Defenders Conference. Nov. 2025.

T11. *Building Effective AI-Powered Data Systems* UMass Amherst Database Seminar. Dec. 2025.

T12. *Building Effective AI-Powered Data Systems* Google BigQuery Team. Nov. 2025.

T13. *Building Effective AI-Powered Data Systems* UCSD Data Systems Seminar. Nov. 2025.

T14. *Building Effective AI-Powered Data Systems* Purdue University Database Seminar. Nov. 2025.

T15. *Building Effective AI-Powered Data Systems* Adobe Research. July 2025.

T16. *Building Effective AI-Powered Data Systems* Redis Labs. June 2025.

T17. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* UC Berkeley BLISS Lab Seminar. Mar. 2025.

T18. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Brown University Data Systems Seminar. Mar. 2025.

T19. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Columbia University Data Systems Seminar. Feb. 2025.

T20. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Scottish Climate Intelligence Service. Feb. 2025.

T21. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Cloudera. Jan. 2025.

T22. *DocWrangler: Steering Semantic Data Processing* LangChain Disrupt Conference. May 2025.

T23. *DocWrangler: Steering Semantic Data Processing* San Francisco Public Defender's Office. Apr. 2025.

T24. *DocWrangler: Steering Semantic Data Processing* Montreal HCI Seminar. Mar. 2025.

T25. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Microsoft Gray Systems Lab. Dec. 2024.

T26. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Snowflake. Nov. 2024.

T27. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* ByteDance (TikTok). Nov. 2024.

T28. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Google Systems Research Group. Nov. 2024.

T29. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* WInE Lab at Carnegie Mellon University. Nov. 2024.

T30. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* Solventum. Nov. 2024.

T31. *DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing* US Army Research Laboratory. Oct. 2024.

T32. *SPADE and EvalGen: Data Quality for LLM Pipelines* HCI Seminar, Hong Kong University of Science and Technology. Aug. 2024.

T33. *SPADE and EvalGen: Data Quality for LLM Pipelines* Databricks Research Seminar. July 2024.

T34. *SPADE and EvalGen: Data Quality for LLM Pipelines* Stanford MLSys Seminar. Apr. 2024.

T35. *SPADE and EvalGen: Data Quality for LLM Pipelines* ML in Practice Seminar, Carnegie Mellon University. Apr. 2024.

T36. *SPADE and EvalGen: Data Quality for LLM Pipelines* Princeton Workshop on Useful and Reliable Agents. Aug. 2024.

Selected Media Coverage

2025 **Lenny's Newsletter (1.1 million subscribers)**, "Building Eval Systems That Improve," Link

2025 **O'Reilly Radar (2.8 million subscribers)**, "Generative AI in the Real World — Shreya Shankar on AI for Corporate Data Processing," Link

2025 **Paper Prisons Initiative, UC Berkeley Criminal Justice Center**, "Using New Generative AI Data Processing Tool DocETL: A Simple Guide," Link

2025 **TWIML AI Podcast (27k subscribers)**, "AI Agents for Data Analysis," Link

2024 **Smol AI News (80k subscribers)**, "DocETL, Agentic Query Rewriting, and the Future of Data Processing," Link

2023 **Weights & Biases Gradient Dissent (62k Subscribers)**, "Operationalizing Machine Learning," Link

2022 **Scale AI Exchange (226k Subscribers)**, "Towards Observability for Machine Learning Pipelines," Link

Teaching

2025–present Co-Instructor and Co-Author, **AI Evals for Engineers and PMs** (professional course and forthcoming O'Reilly book). Co-founded and co-teach with Hamel Husain. Taken by 3,000+ professionals from 500+ companies (50+ each from Google, Microsoft, OpenAI, Meta, Amazon, Intuit, and First American); course reader mailing list exceeds 25,000.

2022 Graduate Student Instructor (TA), **Data 101: Data Engineering**, University of California, Berkeley

2019 Teaching Assistant, **CS101: Computing for Non-Majors**, Stanford University

2016–2018 Teaching Assistant, **CS106B: Programming Abstractions**, Stanford University

Mentorship

Graduate (M.S.) Researchers

2025– **Ruiqi Chen**, University of Washington

2024–2025 **Ankush Garg**, UC Berkeley; now Senior Data Scientist at Clarkston Consulting

2023–2025 **Rachel Lin**, UC Berkeley; now Software Engineer at Opto Investments

Undergraduate Researchers (UC Berkeley unless noted)

2025– **Andrew Cheng**
Harrison Huang
Siyona Sarma
Sasha Singh
Lindsey Wei, University of Washington; nominated for CRA Undergraduate Research Award

2024–2025 **Quentin Romero Lauro**, University of Pittsburgh; nominated for CRA Undergraduate Research Award

2023–2025 **Reya Vir**; now Ph.D. Student at Columbia University; recipient of NSF Graduate Research Fellowship (NSF GRFP)

2022–2024 **Parth Asawa**; now Ph.D. Student at UC Berkeley; CRA Undergraduate Research Award Honorable Mention

2021–2023 **Yujie Wang**; now Software Engineer at Google

2021–2022 **Boyuan Deng**; founded YC-backed startup

2021–2022 **Aditi Mahajan**; now Software Engineer at Google

High School Researchers

2025–2026 **Nikhil Rao** and **Vinay Rao**, Dougherty Valley High School

Academic Service

2027– Reviewer, **Conference on Innovative Data Systems Research (CIDR)**

2026– Reviewer, **Very Large Data Bases (VLDB)**

2024– Reviewer, **ACM Conference on Human Factors in Computing Systems (CHI)**

2025 Organizer, **LLM-Powered Data Processing Workshop**, UC Berkeley — organized for approximately 50 participants (May 2025)

2024– Reviewer, **ACM Symposium on User Interface Software and Technology (UIST)**

2023–2025 Organizer, **DEEM: Data Management for End-to-End Machine Learning Workshop** at SIGMOD

Other Experience

2020–present Advisor and angel investor in early-stage AI startups, including **Alta AI**, **Modal Labs**, **Wispr AI**, **LangChain**, and **Convergent AI**.

2020–present Independent consultant for AI and data infrastructure companies, including **Weights & Biases** and **Parlance Labs**.

2019–2021 **Machine Learning Engineer**, Viaduct (Palo Alto, CA, USA)
First ML engineer; infrastructure for large-scale batch time-series processing; acquired for $104M USD.

Last updated: July 30, 2026