
MUHAMMAD ABU BAKAR SIDDIQUE

Assistant Professor, Department of Computer Science
University of Kentucky, Lexington, KY.

✉ ab.siddique@uky.edu
🌐 cs.uky.edu/~siddique
🎓 Google Scholar

\$4.60M

TOTAL EXTERNAL FUNDING

\$2.00M+

INDIVIDUAL SHARE IN 7 AWARDS

4

DOCTORAL GRADUATES

2

NEW COURSES + B.S. AI MAJOR

RESEARCH PROGRAM

- **Actionable interpretability.** Behavioral testing of large language models can detect undesired behaviors such as sycophancy, reward hacking, and jailbreaks. It cannot explain why they occur, forecast them, or reliably repair them. My research seeks to recover the internal features and circuits that produce such behaviors, causally validate them, and steer the model before the behavior manifests in the output, with minimal side effects on general performance. Our methods have applications in model control, AI safety, and model alignment, among others.
- **Interdisciplinary research.** I work with experts in diverse areas, such as software engineering, cyber-physical systems, cybersecurity, quantum computing, and under-resourced languages. These domains present novel issues such as label scarcity, distribution shift, data leakage, spurious correlations, and unverifiable outputs, which my research aims to address by understanding internal model mechanisms and drawing on insights from domain experts.

EDUCATION

- **University of California, Riverside.** Ph.D., Computer Science *2021*
- **Lahore University of Management Sciences, Pakistan.** M.S., Computer Science *2011*
- **International Islamic University, Islamabad, Pakistan.** B.S., Software Engineering *2008*

APPOINTMENTS

- **University of Kentucky.** Assistant Professor, Computer Science *2021 – present*
- **University of California, Riverside.** Graduate Student Researcher *2017 – 2021*
- **University of Management & Technology, Lahore, Pakistan.** Lecturer *2009 – 2017*

EXTERNAL FUNDING

PI or Co-PI on seven funded awards totaling **\$4.60M**. The institutional share of the University of Kentucky is **\$3.18M**, of which more than **\$2.00M** is my own share. Five of the seven are collaborative, multi-institution awards.

NSF (SHF)	Collaborative Research: SHF: Documentation-Guided Adaptable	\$900,000
Future CoRe	Static Analysis	
	PI. Institutional share \$300,000. 10/2026–09/2029. With Louisiana State University (U. Farooq) and the University of California, Riverside (M. Sridharan).	

NSF (DRL) RITEL	Collaborative Research: RITEL: An Explainable AI Environment for Adaptive Learning and Practice of Quantum Computing PI, with Co-PI K. Huang. Institutional share \$535,000. 08/2026–07/2029. With Louisiana State University (U. Farooq).	\$900,000
SCHMIDT SCIENCES	Medieval Judicial Opinions: Access and Analysis with AI Co-PI, with PI A. Firey. 01/2026–01/2029.	\$400,000
NSF (SaTC) SaTC: CORE	Collaborative Research: SaTC: CORE: Small: Toward Personalized and Trustworthy Mobile App Recommendations via Grounding App Behaviors PI. Institutional share \$300,000. 10/2025–09/2028. With Louisiana State University (U. Farooq).	\$599,994
NSF (CPS) CPS Medium Current: Future CoRe	Collaborative Research: CPS: Medium: An Explainable AI Framework for the Early Detection of Pneumonia in Dairy Calves through Precision Technologies Co-PI, with PI S. Silvestri and Co-PI M. Cantor. Institutional share \$941,359. 09/2025–08/2028. With Pennsylvania State University and the University of Delaware (M. Segovia).	\$1,100,000
NSF (CNS) CISE Large (Planning)	Planning: DCL EPSCoR: CISE Large: Distributed Edge Intelligence in Support of Next-Generation Applications Co-PI, with PI S. Silvestri and Co-PI D. Ionel. 10/2024–09/2026. With Pennsylvania State University (T. La Porta, T. He).	\$100,000
NSF (IIS) CISE CORE	III: Small: Toward User-Centric Task-Oriented Dialog Systems via Zero-shot Adaptation and Personalization PI, with Co-PI B. Harrison. 09/2024–08/2027.	\$599,898

SELECTED PUBLICATIONS

Bold names are students I advise or have advised; my own name is underlined. The full publication list is at [Google Scholar](#) or [my webpage](#).

TMLR '26	Haider MU , Rizwan H, Sajjad H, Ju P, <u>Siddique AB</u> <i>Neurons Speak in Ranges: Breaking Free from Discrete Neuronal Attribution</i> Transactions on Machine Learning Research, 2026
NEURIPS '26	Dhruva J, Raza Y , <u>Siddique AB</u> , Silvestri S <i>RESCAST-100K: A Comprehensive Dataset for Cross-Domain Residential Load and Indoor Temperature Forecasting</i> Conference on Neural Information Processing Systems, 2026 (to appear)
COLM '26	Haider MU , Rizwan H, Sajjad H, <u>Siddique AB</u> <i>Multi-Granular Node Pruning for Causal Circuit Discovery</i> Conference on Language Modeling, 2026 (to appear)

- COLM '26** **Wong SH, Sajjad H, Siddique AB**
LangFIR: Discovering Sparse Language-Specific Features from Monolingual Data for Language Steering
 Conference on Language Modeling, 2026 (to appear)
- EACL '26** **Fereidouni M, Haider MU, Ju P, Siddique AB**
Evaluating Sparse Autoencoders for Monosemantic Representation
 Findings of the Association for Computational Linguistics: EACL, 2026, pp. 5969–5984
- ICSC '26** **Haider MU, Farooq U, Siddique AB, Marron M**
Looking into Black Box Code Language Models
 IEEE International Conference on Semantic Computing, 2026, pp. 32–39
- BEST PAPER NOMINEE**
S&P '25 **Strauss J, Upadhyay K, Siddique AB, Baggili I, Farooq U**
Assessing and Enhancing Quantum Readiness in Mobile Apps
 IEEE Symposium on Security and Privacy, 2025
- QSW '25** **Upadhyay K, Chhetri V, Siddique AB, Farooq U**
Analyzing the Evolution and Maintenance of Quantum Software Repositories
 IEEE International Conference on Quantum Software, 2025
- CustomNLP4U '24** **Fereidouni M, Mosharrof A, Siddique AB**
Grounded Language Agent for Product Search via Intelligent Web Interactions
 1st Workshop on Customizable NLP at EMNLP, 2024, pp. 63–75
- SIGIR '23** **Maqbool MH, Farooq U, Mosharrof A, Siddique AB, Foroosh H**
MobileRec: A Large Scale Dataset for Mobile Apps Recommendation
 46th ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 3007–3016
- WWW '23** **Mosharrof A, Fereidouni M, Siddique AB**
Toward Open-Domain Slot Filling via Self-Supervised Co-Training
 The ACM Web Conference, 2023, pp. 1928–1937. Acceptance rate 19.2%
- ACSAC '23** **Dina AS, Siddique AB, Manivannan D**
FS3: Few-Shot and Self-Supervised Framework for Efficient Intrusion Detection in Internet of Things Networks
 39th Annual Computer Security Applications Conference, 2023, pp. 138–149. Acceptance rate 23.32%
- CIKM '22** **Siddique AB, Maqbool MH, Taywade K, Foroosh H**
Personalizing Task-Oriented Dialog Systems via Zero-Shot Generalizable Reward Function
 31st ACM International Conference on Information and Knowledge Management, 2022, pp. 1787–1797. Acceptance rate 27.5%
- SIGIR '21** **Siddique AB, Jamour F, Xu L, Hristidis V**
Generalized Zero-Shot Intent Detection via Commonsense Knowledge
 44th ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1925–1929. Acceptance rate 27.6%
- WWW '21** **Siddique AB, Jamour F, Hristidis V**
Linguistically-Enriched and Context-Aware Zero-Shot Slot Filling
 The Web Conference, 2021, pp. 3279–3290. Acceptance rate 20.6%

- ICSC '21 Le N, Siddique AB, Jamour F, Oymak S, Hristidis V
 BEST STUDENT PAPER *Predictable and Adaptive Goal-Oriented Dialog Policy Generation*
 IEEE International Conference on Semantic Computing, 2021, pp. 40–47. Acceptance rate 14.9%
- KDD '20 Siddique AB, Oymak S, Hristidis V
Unsupervised Paraphrasing via Deep Reinforcement Learning
 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 1800–1809. Acceptance rate 16.8%
- VLDB '19 Siddique AB, Eldawy A, Hristidis V
Comparing Synopsis Techniques for Approximate Spatial Data Analysis
 45th International Conference on Very Large Data Bases, 2019. Acceptance rate 18.9%

TEACHING AND CURRICULUM DEVELOPMENT

NEW COURSE DEVELOPMENT AND APPROVAL

- **CS 301: Practical Programming for Artificial Intelligence.** Developed and secured approval for a new undergraduate course on the development and deployment of AI systems, covering programming foundations, work with tabular, image, text, and audio data, feature extraction and visualization, and model training, evaluation, and deployment, with fairness and security treated throughout. First offered in Fall 2026.
- **CS 566: Natural Language Processing.** Developed and secured approval for a new graduate course on the theoretical foundations and practical implementation of NLP systems, covering text classification, information extraction, question answering, summarization, machine translation, and conversational systems, with domain applications in finance, law, and e-commerce and a unit on bias, fairness, and societal impact.

CURRICULUM DEVELOPMENT

- Committee member for the university's new B.S. degree in Artificial Intelligence *2024 – 2025*
- Committee member for an Artificial Intelligence Certificate *2024 – 2025*
- Committee member for a proposed Ph.D. program in Bioinformatics *2023 – 2024*

COURSES TAUGHT AT THE UNIVERSITY OF KENTUCKY

- CS 301: Practical Programming for Artificial Intelligence *Fall 2026*
- CS 460G: Machine Learning *Fall 2021–2025*
- CS 515: Algorithm Design *Fall 2023–2026*
- CS 505: Intermediate Topics in Database Systems *Spring 2024–2026*
- CS 585/685: Special Topics, NLP with Deep Neural Networks *Spring 2022, 2023*
- **Earlier teaching.** Sixteen distinct undergraduate and graduate courses taught as a lecturer at the University of Management & Technology (Lahore, Pakistan) between 2009 and 2017, including artificial intelligence, analysis of algorithms, software engineering, big data programming, and mobile application development.

ADVISING AND MENTORING

Four doctoral students have completed their degrees under my supervision or co-supervision. I am the primary advisor for seven doctoral students now in progress, the first of them expected to graduate in 2027.

PH.D. 2026	Adib Mosharrof joined Amazon Advancing Task-Oriented Dialog Systems: Scalability, Generalization, and Evaluation	Primary advisor
PH.D. 2025	Oscar Skea joined Stripe Information-Theoretic Methods for Efficient Training and Robust Evaluations in Self-Supervised Learning	Co-advisor
PH.D. 2025	Muhammad Hasan Maqbool University of Central Florida; joined NVIDIA From Intent Detection to Recommendations: Leveraging Pre-Trained Language Models for Enhanced User Experiences	Co-advisor
PH.D. 2023	Ayesha Siddiqua Dina joined Florida Polytechnic University Deep Learning-Based Intrusion Detection System in Computer Networks and Privacy-Preserving Message Dissemination in Vehicular Ad Hoc Networks	Co-advisor
MASTER'S	Five master's projects supervised to completion, 2024 to 2025, on conversational recommendation datasets, schema-guided dialogue, medieval Latin morphology, biomedical feature selection, and web environments for evaluating autonomous agents. • Halima Abukadah first-authored a paper at IEEE ICSC 2024. • Srijata Maji co-authored a paper at IEEE ICSC 2026.	
UNDERGRAD	Four undergraduate mentees since 2023. • Sing Hieng Wong first-authored a paper at COLM 2026 and has since joined my group as a doctoral student. • Bernardo Alves Rodrigues co-authored a peer-reviewed workshop paper on activation steering. • Katelyn A. Bell holds an REU appointment on the NSF III award.	

DOCTORAL DISSERTATION COMMITTEES

CHAIR OR CO-CHAIR	Chair or co-chair on four completed dissertations and on seven doctoral committees in progress.
MEMBER	Member or outside examiner on nine doctoral committees in progress and on six completed defenses, in the departments of computer science, electrical engineering, mathematics, and physics.

SELECTED INVITED TALKS

- **Keynote.** *N-Grams to ChatGPT: A Journey Through the Evolution of Language Modeling.* 36th International Workshop on Languages and Compilers for Parallel Computing (LCPC). 2023
- *Toward Controllable AI: Understanding and Steering the Behavior of Large Language Models.* AI Literacy and Training Symposium, University of Kentucky. 2026

- *How Do Large Language Models Work? An Explanation of ChatGPT-Like Models and Their Potential for Farms.* Milking the Data Workshop, Pennsylvania State University. 2025
- *Recommender Systems: Bridging Gaps in Data, User Behavior, and Conversations.* Computer Science Colloquium, University of Kentucky. 2025
- *Toward Zero-Shot Generalizable and Personalizable Task-Oriented Dialog Systems.* The Home Depot. 2023
- *Zero-Shot Generalizable Goal-Oriented Dialog Systems.* AI in Medicine seminar, University of Kentucky. 2022

HONORS AND AWARDS

- **Best Paper Award**, IEEE International Conference on Quantum Software 2025
- **Best Paper Nominee**, IEEE International Conference on Semantic Computing 2025, 2026
- **Best Student Paper Award**, IEEE International Conference on Semantic Computing 2021
- **Dissertation-Year Graduate Fellowship**, University of California, Riverside 2021
- **Dean's Fellowship**, University of California, Riverside 2017 – 2018
- **University Tuition Waiver** for the full undergraduate degree 2003 – 2008

PROFESSIONAL SERVICE

- REVIEW PANELS** **National Science Foundation.** Technology, Innovation and Partnerships (TIP), 2023–2025. Information and Intelligent Systems (IIS), 2024–2026. STEM Education, 2026. **Schmidt Sciences**, proposal reviewer, 2026.
- EDITORIAL** Editor, *Big Data Research*, 2023–2026. Reviewer, *ACM Transactions on the Web*.
- PROGRAM COMMITTEES** ICML 2020–2026. NeurIPS 2021–2026. ICLR 2021–2026. COLM 2026. ACL Rolling Review 2021–2026. AAAI 2022–2025. KDD 2023–2025. EMNLP 2022–2024. ACL 2023–2024. NAACL 2024. WWW 2021–2024. SIGIR 2023–2024. CIKM 2022–2024. SDM 2023. IEEE ICSC 2022–2026.
- STUDENT WORKSHOPS** Student Research Workshop program committees: ACL-IJCNLP 2021, NAACL 2021, EACL 2021, ACL 2019–2020, AACL 2020.

INSTITUTIONAL SERVICE AND OUTREACH

- **Selected Institutional Committees**
 - Faculty Search Committee, Department of Computer Science 2024 – 2025
 - Graduate Committee, Department of Computer Science 2021 – 2023
 - Curriculum committees for the B.S. in Artificial Intelligence, an AI Certificate, and a proposed Ph.D. in Bioinformatics 2023 – 2025
- **Outreach**
 - Mentor, Math, Science, and Technology Center, Paul Laurence Dunbar High School, Lexington 2021 – 2025
 - Host a high school student in the lab each summer through the fully funded Summer Youth Program 2022 – 2025