Richard Shan

Generative AI Researcher, The MITRE Corporation

m@richardshan.com
richard-shan
richmshan
richshan

Research Interests

Generative AI, Large Language Models, Retrieval-Augmented Generation (RAG), Mechanistic Interpretability, Machine Learning, Deep Learning, Natural Language Processing, AI Assurance, Computational Linguistics

Research Experience

Present **Generative AI Researcher**, *The MITRE Corporation*, McLean, VA
- Conducted mechanistic interpretability research producing 6 novel metrics to evaluate reasoning performance of LLMs
- Discovered novel phenomenon of feature fracturing and specialization under higher-difficulty tasks
- Identified distinct LLM preferred reasoning "modes" and their feature controls
- Presented findings to Congressional staff about performance of reasoning LLMs in cyberspace

Present **AI Alignment Research Intern**, *Stanford University AIMI Center*, Stanford, CA
- Led team to develop pneumonia chest x-ray detection model with augmentations from NLP on clinical summaries
- Won 2nd place at Stanford-internal hackathon
- Continued work on Automated Bone Fracture Detection System with Integrative Biomedical Imaging Informatics lab

2023 - 2024 **Machine Learning & Software Engineer**, *Contention AI*, Remote
- Built world's first AI-powered debate search engine using machine learning
- Developed backend features for AI large language search engine indexing millions of scholarly articles
- Built machine learning algorithm for article recommendation based on user interactions
- Developed algorithm to highlight important quotations from literature works

2022 **Researcher**, *Wake Forest University Institute for Regenerative Medicine*, Winston-Salem, NC
- Investigated suitability of different 3D printers for telemedicine use cases
- Researched regenerative engineering applications in healthcare
- Presented findings to panel of industry experts

Publications

Peer-Reviewed Journal Articles

2025 R. Shan, "AI That Learns, Thinks, and Acts: The Next Frontier of Generative AI," *IEEE Computer*, vol. 58, no. 10, pp. 28–39, Oct. 2025.

2024 R. Shan, "Certifying Generative AI: Retrieval-Augmented Generation Chatbots in High-Stakes Environments," *IEEE Computer*, vol. 57, no. 9, pp. 32–44, Sept. 2024. Cover Feature.

2024 R. Shan, "Language Artificial Intelligence at a Crossroads: Deciphering the Future of Small and Large Language Models," *IEEE Computer*, vol. 57, no. 8, pp. 28–35, Aug. 2024. Cover Feature, #1 Trending Article.

Peer-Reviewed Conference Proceedings

2025 R. Shan, "LearnRAG: Implementing Retrieval-Augmented Generation for Adaptive Learning Systems," in *2025 IEEE International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2025.

2024 R. Shan et al., "Retrieval-Augmented Generation Architecture Framework: Harnessing the Power of RAG," in *2024 International Conference on Cognitive Computing*, Springer, pp. 86–104, 2024. doi:10.1007/978-3-031-77954-1_6

2024 R. Shan, "ActionFusion: Framework of Large Action Model Enablement," in *2024 IEEE 12th International Conference on Big Data (BigData)*, pp. 9432–9441, 2024.

2024 R. Shan, "A Deep Dive into Vector Stores: Classifying the Backbone of Retrieval-Augmented Generation," in *2024 IEEE International Conference on Big Data (BigData)*, pp. 8831–8833, 2024.

2024 R. Shan, "OpenRAG: Open-source Retrieval-Augmented Generation Architecture for Personalized Learning," in *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication*, 2024.

2024 R. Shan, "A Hybrid Ensemble Approach for Depression Detection: Combining Deep Learning and Machine Learning," in *2024 IEEE 4th International Conference on Data Science and Computer Application*, 2024.

2024 R. Shan, "Enterprise LLMOps: Advancing Large Language Models Operations Practice," *IEEE Cloud Summit*, vol. 143–148, June 2024.

2024 R. Shan, "RAG Engineering," *IEEE Cloud Summit*, 2024.

2023 R. Shan et al., "Edge AI Patterns," in *2023 International Conference on AI and Mobile Services*, pp. 102–110, 2023.

2021 R. Shan et al., "Digital Transformation Method for Healthcare Data," in *2021 International Conference on Big Data*, pp. 48–63, 2021.

Conference Presentations & Invited Talks

2025 "Practical Generative AI Observability: Metrics and Tools for Real-Time Monitoring," *All Things Open AI 2025*, Raleigh, NC.

2025 Oral Presentation at *North Carolina Junior Science and Humanities Symposium (JSHS)*, UNC Charlotte, 2025. Presented LLM evaluation research to Department of Defense representatives.

2024 "The Language Model of Tomorrow: Charting the Course of Generative AI ," *TechStrong Con 2024*.

2024 "A Comparative Study of Machine Learning and Deep Learning Models in Depression Detection," *20th State of North Carolina Undergraduate Research & Creativity Symposium*, 2024.

2023 "Multicloud Deployment Patterns," *DeveloperWeek CloudX 2023 Conference*.

2023 "Pragmatic GraphQL Patterns," *APIDays 2023 Conference*.

2023 "Edge-Oriented Data Patterns" *DataOps Day 2023 Conference*.

2023 "Edge Patterns in Multicloud Environment," *IoTSlam 2023 Conference*.

Research Awards & Honors

Research Competition Awards

2025 **Regional Winner**, *North Carolina Science & Engineering Fair (NCSEF)*, Technology Category

2025 **Category Winner**, *North Carolina Student Academy of Science (NCSAS)*, Technology Category

2025 **Oral Presenter**, *Junior Science & Humanities Symposium (JSHS)*, NC Statewide

2024 **Selected Publication**, *Broad Street Scientific Journal*, Computer Science Category

Research Grants & Fellowships

2025 **Bowman-Brockman Endowment Grantee**
Research funding for AI/ML projects

2025 **Anthropic Grantee**
LLM API access for research